Uniwersytet Ekonomiczny w Poznaniu

Katedra Informatyki Ekonomicznej

Praca doktorska

Metoda ekstrakcji modeli wyceny składki ubezpieczeniowej ze źródeł internetowych

Autor

Piotr Stolarski

Promotor: Prof. dr hab. Witold Abramowicz

Promotor pomocniczy: dr Krzysztof Węcel

Poznań 2015

Pragnę podziękować wszystkim osobom, które przyczyniły się do powstania niniejszej pracy. W szczególności składam podziękowania Promotorom, Rodzinie oraz Współpracownikom z Uniwersytetu Ekonomicznego w Poznaniu.

Spis treści

Spis ilustracji

Spis tabel

Spis skrótów i symboli

AIM - Automatic Induction of binary Machine code

AJAX - Asynchronous JavaScript and XML

B2B – Business-to-business

B2C – Business-to-customer

B2E – Business-to-employee

CAPTCHA - Completely Automated Public Turing test to tell Computers and Humans Apart

CART - Classification And Regression Tree

CDWS - całkowity dopuszczalny współczynnik strat

CHAID - CHi-squared Automatic Interaction Detector

CIT - Conditional Inference Trees

CLI – Common Language Infrastructure

CSS - Cascading Style Sheets

CSV - Comma Separated Values

DOM - Document Object Model

DZWS - dopuszczalny zmienny współczynnik strat

EI - Ekstrakcja informacji

GLM – Generalized Linear Model

GUI - Graficzny Interfejs Użytkownika

HTML - HyperText Markup Language

HTTP(S) - Hypertext Transfer Protocol (Secure)

JSON - JavaScript Object Notation

LARS - least-angle regression

MARS – Multivariate Adaptive Regression Spline

MBR - metody pamięciowe

MSE - średni błąd kwadratowy

OECD – Organizacja Współpracy Gospodarczej i Rozwoju

ORM – Object-Relational Mapping

OWL – Web Ontology Language

OWU – Ogólne Warunki Ubezpieczenia

P3P - the Platform for Privacy Preferences

PLS - cząstkowe najmniejsze kwadraty

PG – Programowanie Genetyczne

PoS - Point of Sale

R^2 - współczynnik determinacji

RDF – Resource Description Framework

RSS – Really Simple Syndication

SQL – Structured Query Language

SVG - Scalable Vector Graphics

UML – Unified Modeling Language

URI - Uniform Resource Identifier

URL - Uniform Resource Locator

WS - współczynnik szkodowości

WWW – World Wide Web

VIN – Vehicle Identification Number

XHTML - Extensible HyperText Markup Language

XML - Extensible Markup Language

XPDL - XML Process Definition Language

XSLT - Extensible Stylesheet Language Transformations

1. Wprowadzenie

1.1 Motywacja

Zjawiskiem, z którym mamy do czynienia w gospodarce w skali międzynarodowej, jest migracja działalności biznesowej, a w szczególności handlu, do elektronicznych kanałów sprzedaży. Raport „Europe's eCommerce Forecast: 2006 to 2011" [Favier2006] przewidywał, że sprzedaż netto w latach 2006-11 wzrośnie z 102 miliardów euro do 263 miliardów euro. Jednocześnie w innym raporcie Forester [Mulpuru2011] prognozowane są w okresie 2010-15 wzrosty rzędu 10%, aby osiągnąć wartość 278 miliardów dolarów w USA oraz odpowiednio 11% do wartości 115 miliardów euro w Europie zachodniej[1].

Przewidywany gwałtowny rozwój e-commerce wydaje się być tym bardziej uwiarygodniony, że w dalszym ciągu wzrasta dostęp społeczeństwa do internetu. Jeśli chodzi o uwarunkowania w Polsce, to z raportu firmy NetTrack [NetTrack2015] wynika, że w 2015 roku aktywne korzystanie z internetu zadeklarowało ponad 76,6% Polaków (czyli ponad 23 miliony osób). Dla porównania w 2010 roku po raz pierwszy było to ponad 50% respondentów w grupie wiekowej 15 i więcej lat [NetTrack2011]. Charakterystyczny jest również fakt, że w badaniu z 2010 roku 72% spośród osób korzystających z internetu stwierdziło, iż korzysta z sieci „codziennie lub prawie codziennie"[2]. Tymczasem w 2015 roku na drugim miejscu wśród celów użytkowników internetu znalazło się robienie zakupów. Takie wykorzystanie zadeklarowało w badaniu ponad 80% internaturów [NetTrack2015]. Nowe trendy polegają też na poprawie jakości usług dostępu do sieci oraz drastycznym wzroście penetracji rynku urządzeń mobilnych. I tak, według raportu „OECD Broadband Portal" [OECD2011], w okresie od czerwca do grudnia 2011 roku Polska była na pierwszym miejscu, jeśli chodzi o wzrost dostępu do szerokopasmowego internetu na 100 mieszkańców z wynikiem 5,8%. W tym samym raporcie czytamy, że dostęp do internetu drogą radiową (a więc

[1] Dla porównania, w USA w tym samym czasie wzrost sprzedaży tradycyjnej ma oscylować wokół poziomu 2,5% rocznie.

[2] Co ciekawe, 98,2% internautów używa internetu w domu. Dla porównania w 2010 roku było to 93,8%, zaś wykorzystanie niegdyś bardzo popularnych kawiarni internetowych spadło już wówczas do marginalnego poziomu 0,6%. W cytowanym badaniu z 2015 roku kategoria ta została usunięta i zastąpiona dostępem za pomocą urządzeń mobilnych z wynikiem bliskim 30%.

także przez sieci komórkowe) w Polsce ma penetrację na poziomie 53,5%[3]. Istotnym aspektem związanym z użytkowaniem urządzeń mobilnych jest fakt, że dają one aktualnie równoprawny dostęp do zasobów informacji w sieci co, jeszcze niedawno, tradycyjny sprzęt komputerowy. W zakresie dziedziny ubezpieczeń trend taki przekłada się na pojawiające się - póki co nielicznie - przypadki sprzedaży ubezpieczeń za pomocą aplikacji dedykowanych na urządzenia mobilne[4].

Wirtualizacja kanałów sprzedaży powoduje także zasadnicze zmiany w mechanizmach obiegu pieniądza. Stąd w obszarze finansów i wielu usług pokrewnych zjawisko wykorzystania potencjału internetu jest także bardzo odczuwalne. Szczególnie dotyczy to bankowości, gdzie aktualnie praktycznie każdy duży bank ma swój system bankowości internetowej. Wiele banków oferuje także rozwiązania mobilne. Według raportu Deutsche Bank PBC odsetek osób korzystających w Polsce z takich systemów w 2011 roku wyniósł 25%. W 2012 roku ponad 40% Europejczyków używało bankowości elektronicznej, a na pytanie o najchętniej wybierany kanał kontaktu z bankiem 55,1% respondentów wybrało właśnie bankowość elektroniczną[5].

Dużo ostrożniejsze, do tej pory, jest wykorzystanie nowych technologii zwłaszcza w zakresie dystrybucji w ubezpieczeniach[6]. Niemniej obserwacje rynku ubezpieczeniowego wskazują na trend wzrostowy wolumenu sprzedaży produktów i usług na tym rynku przez internet na świecie[7]. Również dane zebrane przez nas w porównaniu z poprzednimi badaniami potwierdzają rosnącą rolę kanału internetowego jako narzędzia marketingowego oraz dystrybucji w Polsce. W sieci pojawia się coraz większa liczba ubezpieczycieli i pośredników. Prognozy wskazują, że proces ten będzie nadal postępował[8].

Można zatem racjonalnie oczekiwać, że w najbliższych latach konkurencja w zakresie masowych produktów ubezpieczeniowych będzie koncentrowała się na walce

[3] Taki wynik plasuje kraj w pierwszej piętnastce krajów świata. Na zbliżonym poziomie są: Wielka Brytania - 53,5%, Szwajcaria - 53%, Holandia - 52,5%.
[4] W tym przypadku mowa może być zatem o m-commerce.
[5] To o blisko 1,5 p.p. więcej niż wizytę w oddziale.
[6] Porównanie sektora ubezpieczeń oraz bankowości wydaje się być uzasadnione licznymi podobieństwami pomiędzy obydwoma jak również występującymi pomiędzy nimi powiązaniami.
[7] Patrz np. "US Online Insurance Forecast, 2010 To 2015", Forrester Research, Inc. 2011 oraz http://www.bankier.pl/wiadomosc/Co-kilka-sekund-ktos-dzwoni-lub-klika-po-polise-2427449.html, odczytano 12-12-2011.
[8] http://bluemedia.pl/press_room/eurobank_-/, odczytano 03-11-2012.
http://bluemedia.pl/projekty/dla_klientow_zewnetrznych/direct/, odczytano 03-11-2012.

o zdobycie klienta w kanale dystrybucji przez internet. Stąd prowadzenie badań zorientowanych na ten segment rynku wydaje się być szczególnie istotne oraz aktualne.

W pracy posługujemy się pojęciem **wiedza ubezpieczeniowa**, które definiujemy jako każdy zasób wiedzy bezpośrednio dotyczący rynku lub produktu ubezpieczeniowego. Szczególnym przypadkiem takiej wiedzy są modele wyceny produktów ubezpieczeniowych. Określone usługi w internecie - źródła internetowe oferujące ubezpieczenia na sprzedaż - mogą być źródłem nieujawnionej[9] wiedzy ubezpieczeniowej, wartościowej dla osób i podmiotów zajmujących się badaniem rynku ubezpieczeń oraz jego uczestników, w tym samych zakładów ubezpieczeniowych. Należy podkreślić, że pracownicy firm ubezpieczeniowych oraz badacze do tej pory nie zdawali sobie sprawy lub ignorowali możliwości użycia internetu do badania oferty produktowej. Praktyka gospodarcza rynku ubezpieczeń dostarcza licznych przykładów na to, że znajomość sposobu kształtowania ceny produktu jest czynnikiem pożądanym.

Podsumowując opisane powyżej przypadki stwierdzić należy, że wiedza ubezpieczeniowa stanowi użyteczny zasób szczególnie dla celów analizy i badania rynku. Elektroniczne kanały dystrybucji dają możliwość zmniejszenia asymetrii informacji[10] oraz redukcji zjawiska „szumu" informacyjnego[11]. Ponadto wiedza pozyskana z portali i serwisów ubezpieczeniowych może mieć szereg zastosowań – zarówno dla klientów, jak też podmiotów branżowych. Najistotniejsze to:

- badania naukowe związane z porównywaniem modeli i kształtowania cen oraz ich zmianami w czasie,
- nadzór nad sprzedażą ubezpieczeń w internecie oraz audyt[12,13],

[9] W literaturze ekonomicznej poza rozróżnieniem wiedzy jawnej (explicit) i ukrytej (tacit) [Nonaka1995], rozważa się także istnienie wiedzy tajnej (hidden) [Riley1985].

[10] Jest to zatem podejścia zgodne z założeniami gospodarki opartej na wiedzy i umożliwia wspieranie optymalizacji kosztowej decyzji konsumenckich. Tematyka asymetrii informacji poruszana jest w szeregu tekstów m.in. [Dionne1992], [Rotschild1976].

[11] W badaniu opinii internautów przeprowadzonym przez Gemius SA a przygotowanym na zlecenie Alianz wymieniono szereg negatywnych skojarzeń użytkowników z popularnymi ubezpieczycielami. Jednym z częściej pojawiającym się czynników jest niezgodność reklam z rzeczywistą ofertą.

[12] W [Werner2010] problem ten prezentowany jest następująco: "Niektóre państwa wprowadzają regulacje, które zawierają szczegółowe określenie tego, co jest dozwolone oraz niedozwolone w ocenie klasyfikacji ryzyka dla różnych produktów ubezpieczeniowych. Koniecznością jest, aby system klasyfikacji oraz taryfikacji był zgodny z obowiązującymi przepisami ustawowymi i wykonawczymi każdej danej jurysdykcji, w której firma prowadzi działalność. [...] Niektóre państwa mogą zezwolić na użycie określonej zmiennej taryfikacyjnej, jednocześnie nakładając ograniczenia na jej zastosowanie. [...] W innych przypadkach prawodawstwo może zakazać użycia pewnych zmiennych w samym algorytmie kalkulacji stawki, ale zezwalać na ich stosowanie na etapie dopuszczenia ryzyka do ubezpieczenia. Zmienne użyte na etapie oceny dopuszczalności mogą być

- monitorowanie rynku i konkurencji,
- wykorzystanie modeli konkurencji dla wsparcia własnej sprzedaży ubezpieczyciela przez stworzenie punktu odniesienia do porównania oferty z innymi firmami, kosztami ubezpieczeń oferowanych przez agentów[14],
- strategie imitacji i wzorowania przy tworzeniu produktów własnych na etapie projektowania produktu (aktuarialnym)[15],
- re-engineering modelu wyceny produktu już oferowanego, co może zapobiegać efektowi negatywnej selekcji klientów w portfelu[16],
- ustanawianie standardów opisu modeli wyceny produktów, transfer dobrych praktyk i doświadczeń pomiędzy produktami,
- tworzenie meta-modeli (szablonów produktowych) – próby tego rodzaju podejść są rzadkie, ale pojawiły się już w literaturze ubezpieczeniowej [Dionne1989],
- ułatwienie interoperacyjności w zakresie wymiany wiedzy – szczególnie istotne dla portali i systemów porównujących oferty i pośredników. Systemy takie mogą nie tylko w łatwy sposób posiąść wiedzę o nowych produk-

---

stosowane do podejmowania decyzji w zakresie selekcji ryzyk, ale mogą także służyć do kształtowania polityki w zakresie pozycjonowania i budowy struktury portfela ryzyk."

[13] Ciekawym przypadkiem biznesowym może być tutaj chociażby kontrola oraz ocena wpływu wymogu implementacji regulacji Komisji Europejskiej zakazującej dyskryminacji ze względu na płeć przy obliczaniu poziomu składki - http://samcik.blox.pl/2012/11/Allianz-straszy-kobiety-koncem-swiata-Mozna-sie.html, odczytano 25-11-2012.

[14] Jest to istotny aspekt, którego brak wymieniany jest przez internautów we wspomnianym już raporcie Gemius SA dla Alianz.

[15] Autorzy cytowanej pozycji [Werner2010] piszą w niej m.in.: „Firmy opracowujące zupełnie nowy produkt ubezpieczeniowy zazwyczaj nie mają danych niezbędnych do projektowania składników podstawowego równania ubezpieczenia. W związku z tym, firmy te na ogół opierają się na informacjach z ich innych podobnych produktów, podobnych produktów oferowanych przez konkurentów lub informacji z biur ratingowych, dokonując odpowiednich korekt. [...] firma może użyć instrukcję konkurenta jako punkt wyjścia i dostosować na podstawie różnic znanych lub oczekiwanych. Dodatkowo do podręcznika sprzedaży polis konkurenta, firma powinna starać się uzyskać informacje na temat względnych poziomów wydatków i rentowności konkurenta. [...] Firma może wykorzystać te informacje, aby lepiej oszacować oczekiwany zysk, jeśli kopiuje stawki konkurenta."

[16] Efekt ten opisywany jest wielokrotnie w literaturze. Ze względu na powiązania z mechanizmami opracowywania stawek, piszą o nim również obszernie autorzy w [Werner2010] m.in. w taki sposób: "Wydaje się, że przedsiębiorstwo powinno być zadowolone, tak długo jak poziom stawek gwarantuje uzyskanie żądanego zagregowanego zysku i nie powinno być zatem zbyt zainteresowane w indywidualnej adekwatności składek. W rzeczywistości, firma, która nie pobiera właściwej stawki dla poszczególnych ryzyk, podczas gdy inne firmy pobierają, może być poddana negatywnej selekcji, co w konsekwencji wpłynie na pogorszenie wyników finansowych. Ponadto firma, która różnicuje ryzyka przy użyciu właściwych charakterystyk, których inne firmy nie stosują, będzie podlegała zjawisku pozytywnej selekcji, w konsekwencji uzyskując przewagę konkurencyjną."

tach, ale także łatwo weryfikować momenty, kiedy produkty te ulegają zmianie.

Odkrywaniu wiedzy ubezpieczeniowej ze źródeł internetowych towarzyszy szereg praktycznych problemów, takich jak: występowanie zjawisk ograniczenia zasobów, zdolność oceny jakości pozyskanej wiedzy, a także jej aktualność. Do zagadnień tych odniesiemy się w ramach prezentowanej metody ekstrakcji.

1.2 Zakres badań i teza pracy

Obszarem badawczym, w obrębie którego praca powstała, jest zagadnienie ekstrakcji wiedzy ze źródeł webowych[17]. Ekstrakcja wiedzy z internetu jest częścią podejścia badawczego określanego w literaturze jako web mining. Należy jednak podkreślić, że postawiony problem badawczy, znajdujący się na pograniczu specyficznej dziedziny analizowanych źródeł internetowych – witryn ubezpieczeniowych – oraz eksploracji treści witryn, ma charakter oryginalnego wkładu, a najbliżej jest mu do problemu generowania skrótów i podsumowań w ramach przedstawionego stanu badań prowadzonego w odnośnym obszarze.

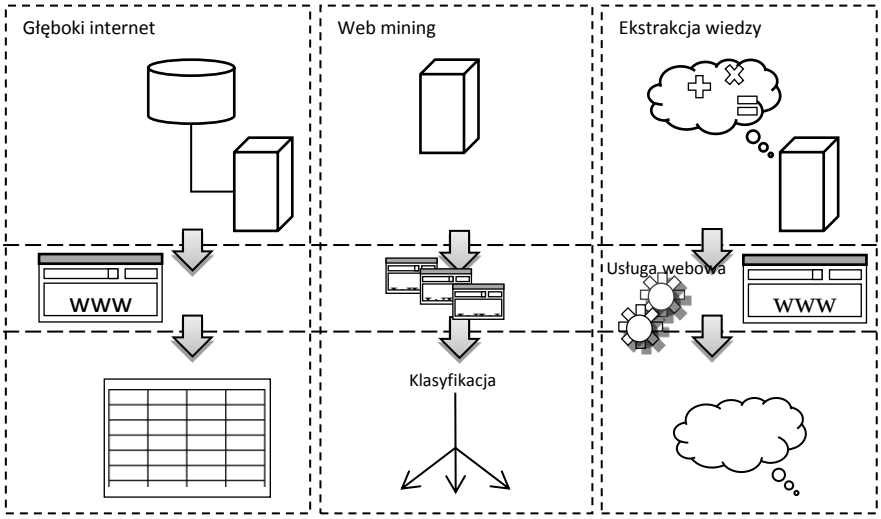
Zakres pracy wykracza jednak poza tematykę ekstrakcji danych z treści webowych, korzystając pomocniczo z dwóch dodatkowych obszarów: metod aktuarialnych składających się na metody wyznaczania cen ryzyka i taryfikacji produktów ubezpieczeniowych[18], a także metod analitycznej obróbki danych pozwalających na agregację wiedzy ze zgromadzonych zbiorów danych[19]. W tym ostatnim obszarze szczególnie istotna dla problematyki pracy jest regresja symboliczna wraz z jej uszczegółowieniem, jakim jest metoda programowania genetycznego.

Problem badawczy niniejszej dysertacji określić można w formie następującego zestawu pytań: czy jest możliwe, z punktu widzenia teoretycznego, oraz w przypadku praktycznej realizacji, z wykorzystaniem jakich środków technicznych, pozyskanie wiedzy o produktach ubezpieczeniowych z internetu? W jaki sposób szczegółowo i z zastosowaniem jakiego metodycznego podejścia można tę wiedzę pozyskiwać?

[17] „Źródło webowe" jest w rozumieniu pracy pojęciem węższym niż „źródło internetowe". Niemniej tam, gdzie nie powoduje to wątpliwości obydwa wyrażenia używane są synonimicznie.
[18] Jest to specyficzna wiedza dziedzinowa, która ma charakter pomocniczy na różnych etapach odtwarzania modelu obliczania składki.
[19] Traktowana raczej instrumentalnie w celu dokonania selekcji najbardziej obiecujących narzędzi analitycznych.

Problemy w rzeczywistym świecie powiązane są jednocześnie z wyzwaniami, których przezwyciężenie stanowi formę rozwiązania problemu – w przypadku prezentowanego badania podstawowym wyzwaniem jest opracowanie metody, stanowiącej odpowiedź na pierwsze pytanie, oraz prototypu, będącego narzędziem do pozyskiwania wiedzy ubezpieczeniowej, stanowiącego odpowiedź na drugie pytanie.

Ekstrakcją modeli wyceny ubezpieczeń ze źródeł internetowych nazwiemy działanie polegające na zbudowaniu reprezentacji źródła oraz charakterystyk ekstrahowanego modelu służących wyznaczeniu zależności pomiędzy wartościami zmiennych niezależnych modelu[20] wynikających ze wspomnianych charakterystyk a wielkością składki ubezpieczeniowej. Naszym celem jest otrzymanie modelu o możliwie małym błędzie, wyliczanym jako różnica między wartościami przewidywanymi a rzeczywistymi. Jednocześnie optymalizujemy proces odtwarzania modelu polegający na minimalizacji liczby zapytań skierowanych do źródła.



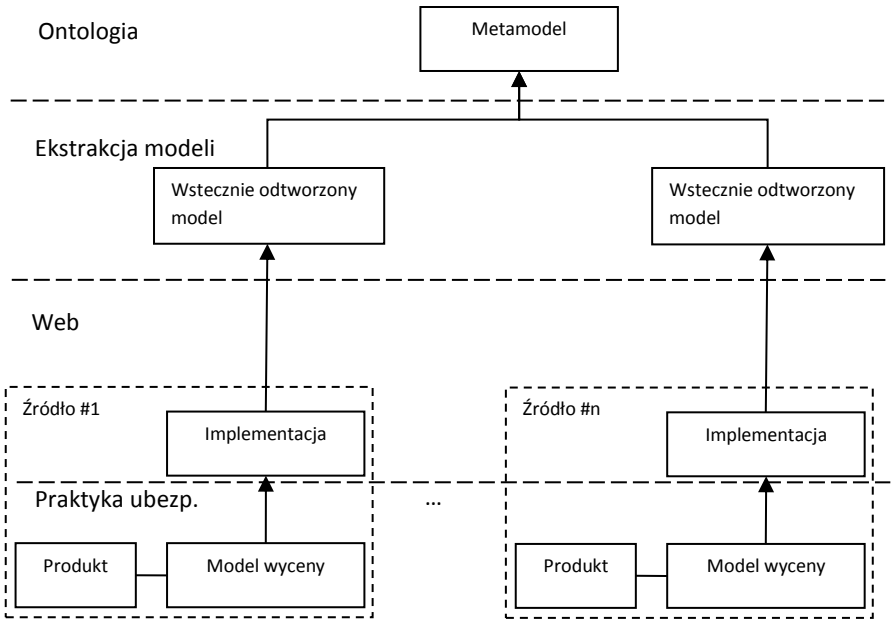
Rysunek 1. Głęboki internet, web mining oraz ekstrakcja wiedzy – porównanie
Źródło: opracowanie własne

Wykorzystując ekstrakcję informacji, nie sięgamy do bazy danych udostępnionej przez stronę internetową[21], lecz staramy się poznać algorytm wyliczania wielkości składki na podstawie zebranych danych. Zatem w odróżnieniu od ekstrakcji informacji, np. ze źródeł głębokiego internetu, w zaproponowanym podejściu zajmujemy się ekstrakcją wiedzy. W pracy dokonujemy szczegółowego przedstawienia metody odtwarzania modelu, omawiamy wyzwania z nią związane oraz przedstawiamy narzę-

[20] Odpowiadających tzw. zmiennym taryfikacyjnym.
[21] Jak ma to miejsce w przypadku ekstrakcji informacji z głębokiego internetu.

dzie wspierające jej użycie. Ponadto pokazujemy i weryfikujemy uzyskane rezultaty, a także dyskutujemy zakres zastosowania.

Problem porównania modeli wyceny wymaga operowania na wyższym poziomie abstrakcji niż poziom technologii ekstrakcji informacji. Naturalnym sposobem agregacji aparatu pojęciowego wydaje się być ontologia dziedzinowa. Przez metamodel rozumiemy tutaj pewien wzorcowy lub uogólniony model wyceny, zawierający np. rozszerzoną listę parametrów lub łączący zależności grup zbliżonych parametrów wraz z ich wpływem na cenę składki [Dionne1989]. Hierarchia modeli wyceny aż do poziomu metamodelu zaprezentowana jest na rysunku 2. Model taki może mieć znaczenie referencyjne jako osobny zasób wiedzy.



Rysunek 2. Poziomy operowania na modelach wyceny
Źródło: opracowanie własne

Celem badania zaprezentowanego w niniejszej dysertacji jest zaproponowanie ogólnej i efektywnej metody ekstrakcji modeli wyceny ubezpieczeń ze źródeł internetowych jako szczególnego przypadku pozyskiwania wiedzy o rynku ubezpieczeń. Wraz z metodą stworzone są szczegółowe algorytmy pozwalające na realizacje wyznaczonego celu badawczego.

W związku z tak wyznaczonym celem badawczym formułujemy następującą tezę pracy:

Opracowana metoda odtworzenia semantycznego taryf, która wykorzystuje semantyczny model dziedziny[22] *produktu ubezpieczeniowego, opracowany mechanizm ekstrakcji danych ze źródeł internetowych oraz odpowiednio dobrane narzędzia odkrywania wiedzy, umożliwia odtwarzanie modeli wyceny składki ubezpieczeniowej.*

Aby osiągnąć opisany cel badawczy, konieczne jest dodatkowo zrealizowanie celów pomocniczych, którymi są:

1. przegląd, kategoryzacja oraz dobór do zbioru badawczego źródeł internetowych będących pożądanymi przypadkami źródeł wiedzy ubezpieczeniowej,
2. rozwój artefaktów niezbędnych lub istotnie poprawiających funkcjonowanie metody ekstrakcji wiedzy, takich jak ontologie czy słowniki pomocnicze,
3. opracowanie poprawnej metodycznie oraz zgodnej z rzeczywistymi możliwościami badawczymi metody ewaluacji i porównania modeli.

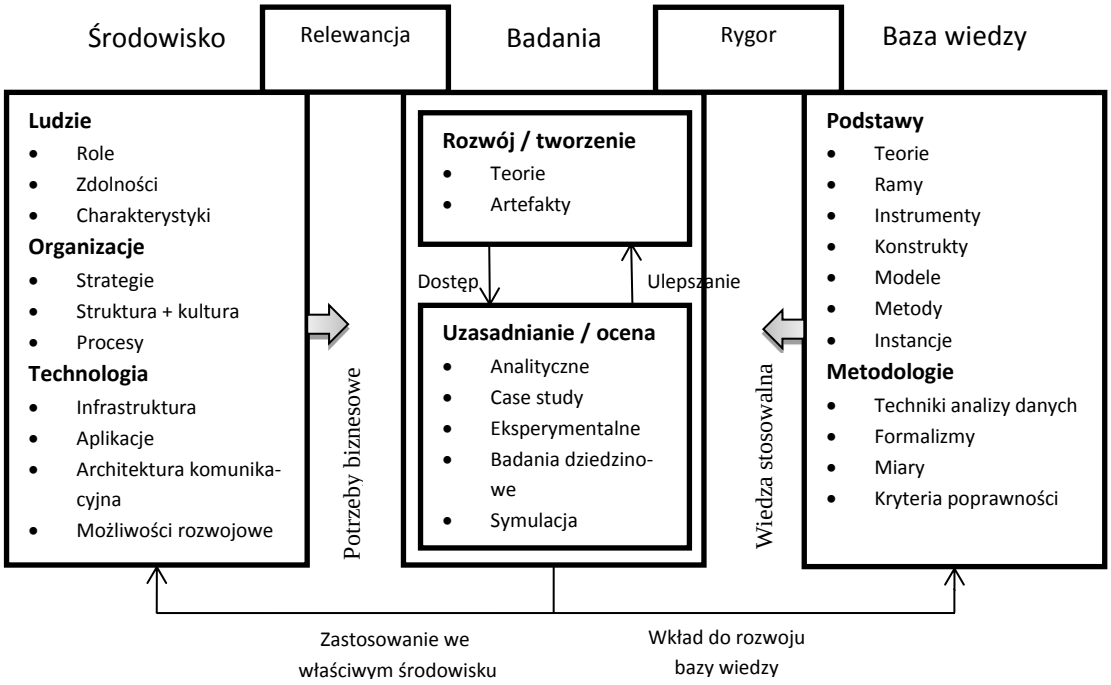
1.3 Metodologia

W dziedzinie badań nad systemami informacyjnymi rozwinęły się zasadniczo dwa istotnie różne paradygmaty uprawiania nauki. Przedmiotem rozważań i analizy w ramach nurtu badań behawioralnych jest zachowanie i zasięg oddziaływania czy też interakcji systemów informacyjnych na środowisko organizacyjne oraz na jednostki w tym środowisku działające. Ze względu na przedstawiony powyżej zakres i program badawczy, paradygmat powyższy nie jest właściwy do zastosowania w kontekście niniejszej dysertacji. Znacznie lepiej dopasowany nurt – badań projektowych[23], za którego sformułowanie odpowiedzialny jest prof. Hevner [Hevner2004] – koncentruje się na koncepcji rozszerzania granic możliwości ludzkich oraz organizacyjnych przez projektowanie i tworzenie nowych artefaktów[24].

Na rysunku 3 przedstawiono schemat koncepcyjny podejścia badawczego proponowanego przez Hevnera. W podejściu tym centralną część zajmują badania polegające na iteracyjnym rozwoju i ocenie różnego rodzaju wytworzonych lub ulepszonych

[22] Dziedzinę – w przypadku ontologii w literaturze przedmiotu przyjęło się mówić o ontologiach domenowych, jako zbiorach (przynajmniej) pojęć i relacji opisujących określoną dziedzinę rzeczywistości.
[23] ang. design science.
[24] W tym sensie podejście to wykazuje pewne podobieństwa do nauk inżynieryjnych

artefaktów. Badania motywowane są potrzebami biznesowymi wywodzonymi z potrzeb lub wskazywanymi przez środowisko. Środowisko rozumiane jest tutaj szeroko: jako zbiory ludzi, organizacje oraz zróżnicowane aspekty technologiczne. Z drugiej strony skuteczne przeprowadzenie prac badawczo-rozwojowych możliwe jest wyłącznie pod warunkiem osadzenia ich w kontekście właściwie przeprowadzonego przeglądu bazy wiedzy. Na bazę wiedzy składają się elementy podstawowe, takie jak: fundamentalne teorie, ramy, istniejące modele i metody etc. oraz czynniki wtórne, jakimi są sposoby ewaluacji.



Rysunek 3. Schemat koncepcyjny podejścia badawczego
Źródło: [Hevner2004]

Prezentowana metodologia dostarcza wreszcie wskazówek umożliwiających identyfikację właściwych wyników badawczych. Są nimi przede wszystkim: modele, metody, instancje[25] oraz inne elementy bazy wiedzy, stanowiące nowy wkład lub istotne ulepszenie stanu obecnego.

W nawiązaniu do powyższego wyszczególnienia, wynikami badawczymi niniejszej pracy są:

[25] O charakterze prototypów lub aplikacji odzwierciedlających wcześniej wytworzone modele i metody.

1. metoda ekstrakcji wiedzy ubezpieczeniowej,
2. wspomagający model wiedzy,
3. prototyp systemu ekstrakcji,
4. instancje modeli przeznaczonych do ewaluacji.

1.4 Struktura pracy

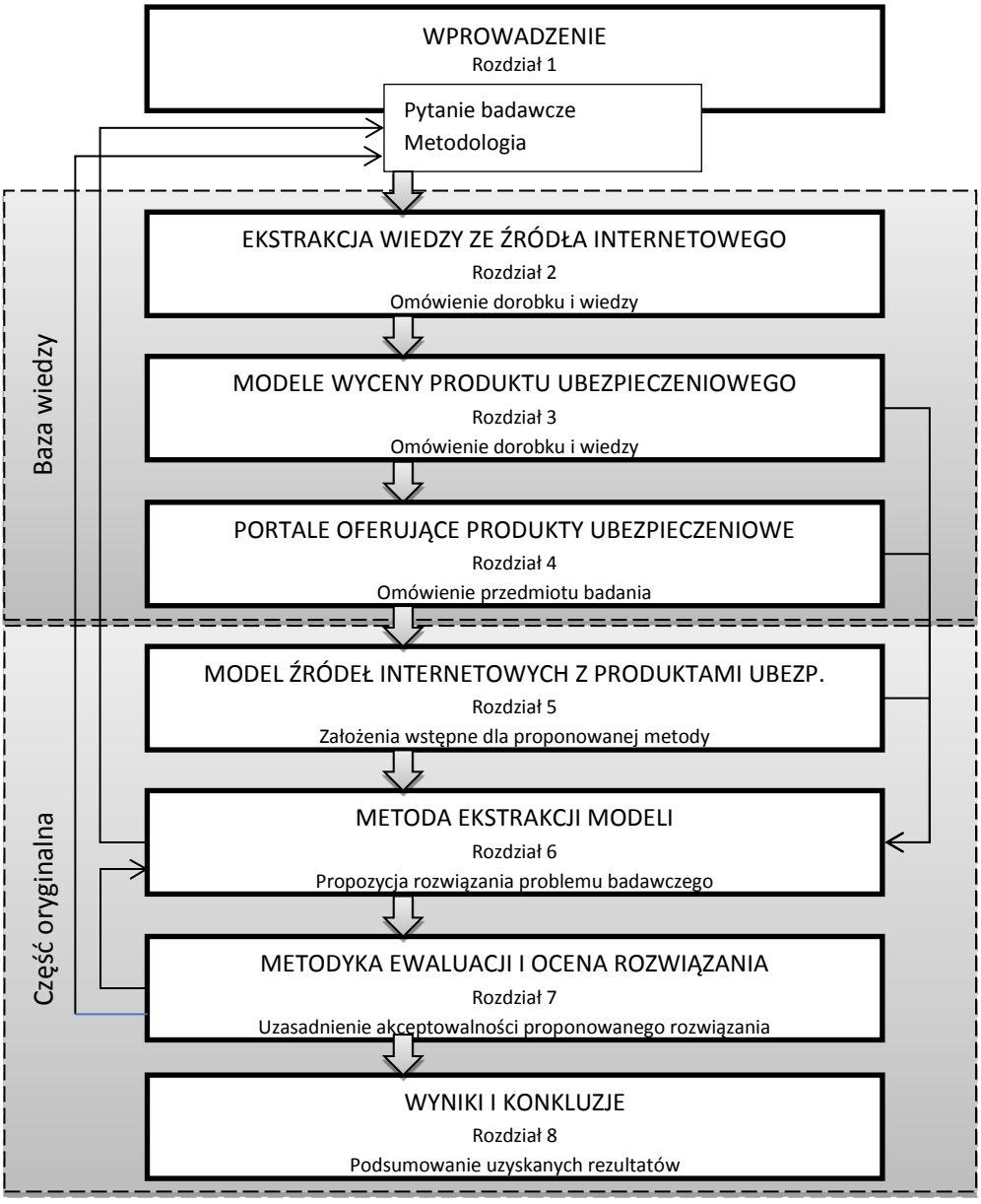
Konsekwentne odniesienie się do problemów badawczych wymaga zrealizowania następujących czynności[26]:

1. przeanalizowanie aktualnej literatury w zakresie ekstrakcji informacji i odkrywania wiedzy ze źródeł webowych,
2. zapoznanie się z technikami i metodami konstrukcji produktów ubezpieczeniowych, wyceny składki oraz ich taryfikacji,
3. kategoryzacja i charakterystyka witryn internetowych w ramach różnorodnych kryteriów dla określenia zakresu i możliwości wykorzystania tworzonej metody,
4. opracowanie metody ekstrakcji wraz z jej wyczerpującym udokumentowaniem,
5. określenie zasad weryfikacji i oceny uzyskanych rezultatów w postaci spójnej metody ewaluacji,
6. przeprowadzenie ewaluacji.

W niniejszej pracy skupiamy się na zadaniu pozyskiwania modeli wyceny produktów ubezpieczeniowych ze źródeł internetowych. Przedstawiony na rysunku 4 schemat pracy odzwierciedla realizację punktów wymienionych powyżej. Jak zademonstrowano na diagramie, praca tradycyjnie wprowadza podział na część referowaną, stanowiącą bazę wiedzy oraz część oryginalną, opisującą badania i wkład do zastanego stanu wiedzy. Na część pierwszą składają się rozdziały 2-4, kolejne rozdziały (5-8) stanowią oryginalne rozwiązanie problemu badawczego, przedstawionego w rozdziale 1 oraz ewaluację wyników. Strzałki na rysunku 4 wskazują istotne powiązania pomiędzy treściami zawartymi w poszczególnych częściach pracy.

[26] Podaną listę czynności można traktować jako program badawczy.

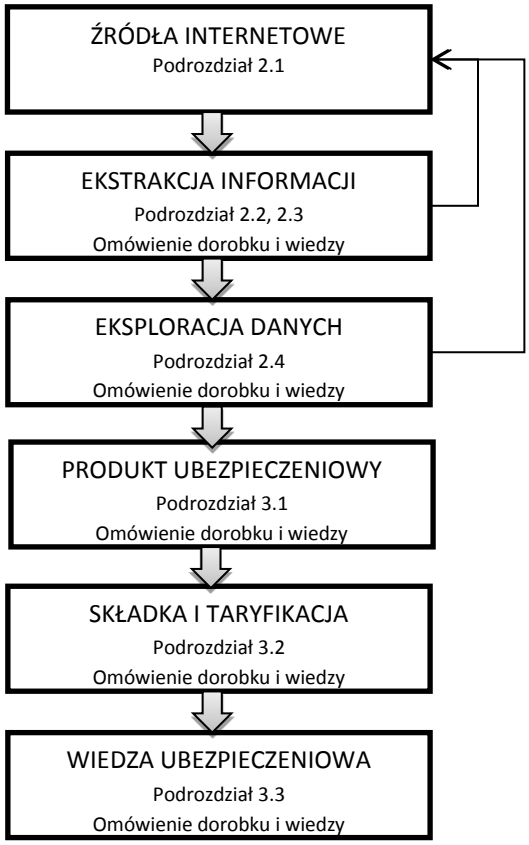
W rozdziale 2 dokonujemy przeglądu prac i rezultatów związanych z zagadnieniami ekstrakcji informacji oraz eksploracji danych ze źródeł webowych, czyli zagadnieniami wykazującymi podobieństwo do przypadku będącego przedmiotem zainteresowania. W rozdziale 3 prezentujemy istotne wiadomości dotyczące problematyki wyceny ubezpieczeń. Rozdział 4 koncentruje się z kolei na zagadnieniach sprzedaży ubezpieczeń przez internet. W rozdziałach 5 i 6 szczegółowo przedstawiamy opis prezentowanej metody, zastosowane ramy teoretyczne oraz ich praktyczne implementacje. W końcu rozdziały 7 i 8 prezentują ostatecznie otrzymane wyniki oraz dyskusję nad użytecznością metody.



Rysunek 4. Schemat pracy
Źródło: opracowanie własne

2 Ekstrakcja wiedzy ze źródła internetowego

W rozdziale tym sukcesywnie prezentujemy stan wiedzy związany z obszarami badawczymi. Prezentacja zgromadzonej na potrzeby dysertacji bazy wiedzy, reprezentującej prawą stronę diagramu znajdującego się na rysunku 3, dokonana zostanie zgodnie z poniższym schematem.



Rysunek 5. Schemat przeglądu prac
Źródło: opracowanie własne

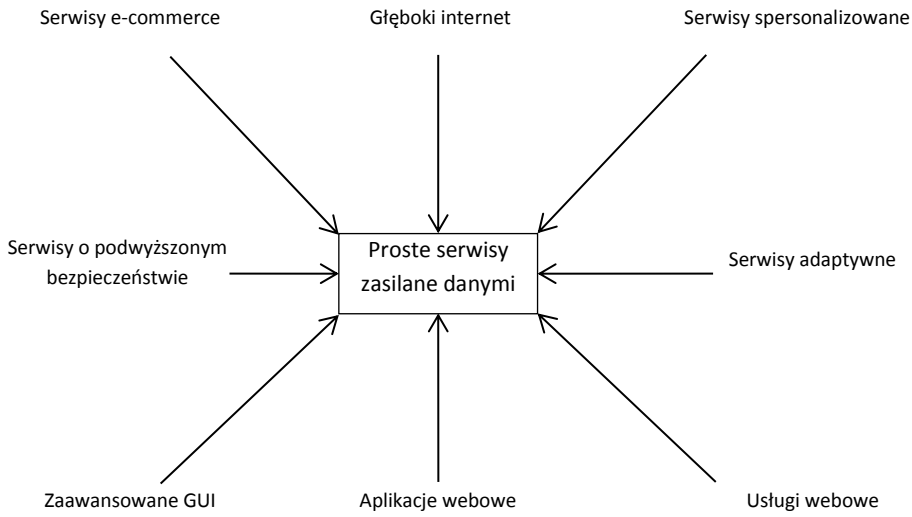
2.1 Źródła internetowe, cechy, klasyfikacja

Źródłem internetowym w rozumieniu pracy jest każdy zasób informacji dostępny za pomocą sieci internet - w szczególności dostępny za pomocą protokołu HTTP(S)[27] (strony WWW, usługi webowe).

Przeprowadzając przegląd literaturowy w zakresie klasyfikacji oraz charakterystyk źródeł internetowych można dojść do wniosku, że istnieje względna zgoda wśród autorów co do systematyzacji takich źródeł. Ze względu na istotne różnice w sposobie funkcjonowania poszczególnych typów źródeł wyróżnionych w ramach przytaczanej

[27] HyperText Transfer Protocol oraz jego bezpieczna (Secure) odmiana.

klasyfikacji, a często także innych różnic wybiegających poza użytkowanie specyficznych mechanizmów i formalizmów, w przypadku poszczególnych elementów klasyfikacji mówić możemy o modelach źródeł. Przez model źródła rozumiemy uproszczone konstrukcje myślowe będące nośnikami tylko cech istotnych ze względu na podział, abstrahujące od szczegółów. Zaznaczyć od razu należy, iż mając na myśli takie modele źródeł przeprowadza się daleko idącą idealizację. Faktyczne źródła internetowe stanowią bowiem w ogromnej większości różne kombinacje wyróżnionych typów idealnych. Klasyfikację źródeł internetowych traktować należy bardziej w kategoriach wymiarów, za pomocą których opisać można poszczególne realne serwisy internetowe.



Rysunek 6. Klasyfikacja źródeł internetowych
Źródło: opracowanie własne

Przechodząc do opisu samej klasyfikacji, zaczniemy od podstawowego typu źródła internetowego, jakim jest prosty serwis zasilany danymi. Każdy inny model serwisu różnić się będzie pewnymi dodatkowymi cechami w stosunku do tego pierwotnie zdefiniowanego. Różnice te wynikać będą z następujących wymiarów przedstawionych na rysunku 6.

2.1.1 Proste serwisy zasilane danymi

Treści udostępniane w internecie można podzielić na: nieustrukturyzowane oraz ustrukturyzowane. Do pierwszej kategorii zaliczymy czyste dokumenty tekstowe bez

oznaczonej struktury oraz multimedia[28]. Dokumenty ustrukturyzowane wymagają posiadania, poza zawartością samej informacji o treści, także dodatkowej informacji opisującej strukturę dokumentu. Informacja o strukturze zapisana jest w przeważającej części dokumentów za pomocą języka znaczników: HTML[29], XML[30] lub kombinacji obu – XHTML[31].

W odróżnieniu od standardu XML, który nie definiuje bezpośrednio leksykonu znaczników, gwarantując przez to jego uniwersalność, HTML i częściowo XHTML[32] dla danej konkretnej wersji standardu posiadają zamkniętą listę znaczników. Jak każdy język wywodzący się ze specyfikacji XML, również języki (X)HTML pozwalają na traktowanie fragmentów treści dokumentu jako elementów przyporządkowanych do wierzchołków drzewa oznaczonych za pomocą znaczników[33]. Równolegle znacznikom tym przyporządkowana jest określona semantyka – w większości przypadków sprowadzająca się do sposobu prezentacji w przeglądarce WWW. Od momentu wprowadzenia formalizmu kaskadowych arkuszy styli CSS[34], reguły interpretacji znaczników uległy komplikacji – możliwa stała się praktycznie dowolna ich modyfikacja w zakresie warstwy prezentacji. Jednocześnie stała się możliwa interpretowana wizualizacja dowolnych znaczników z przestrzeni całego XML.

2.1.2 Głęboki internet

Aby przedstawić istotę różnicy pomiędzy modelem prostego serwisu zasilanego danymi (płytki internet), a modelem stron głębokiego internetu [Bergman2001], należy przeanalizować mechanizm udostępniania treści przez usługę WWW jako typowy system klient-serwer. W systemie takim klient WWW pośredniczy w przesyłaniu żądań do serwera, które są wynikiem interakcji klienta z użytkownikiem. Natomiast serwer te żądania przetwarza i w rezultacie odpowiada na nie, dostarczając treść.

Głęboki internet różni się od płytkiego internetu w dwóch kluczowych obszarach: przebiegu interakcji z użytkownikiem oraz sposobu przetworzenia żądania. Jeśli cho-

[28] Pliki zawierające dane multimedialne posiadają pewną strukturę. Ma ona jednak zasadniczo inny charakter, odrębny jest także sposób przetwarzania takich danych.
[29] Hypertext Markup Language, http://www.w3.org/TR/html401/, odczytano 20-11-2012 r.
[30] Extensible Markup Language, http://www.w3.org/TR/REC-xml/, odczytano 20-11-2012 r.
[31] Extensible HyperText Markup Language, http://www.w3.org/TR/xhtml1/, odczytano 20-11-2012 r.
[32] Dokument w XHTML mogą być rozszerzane jak każdy dokument XML
[33] Chodzi tutaj o tzw. drzewo DOM – Document Object Model.
[34] Cascading Style Sheets, http://www.w3.org/TR/CSS2/, odczytano 20-11-2012 r.

dzi o specyfikę interakcji klienta WWW, to charakterystyczne w modelu głębokiego internetu jest występowanie żądań sparametryzowanych. Dodatkowo o wartościach parametrów tych żądań w istotnym stopniu decyduje sam użytkownik, któremu w warstwie prezentacji strona WWW dostarcza niezbędnej infrastruktury do decydowania o parametrach żądania. Spoglądając z kolei na zagadnienie od strony serwera, sposób przetworzenia żądania jest istotnie bardziej skomplikowany niż obsługa żądań w modelu prostego serwisu wykorzystującego dane. Parametry żądania są bowiem przekazywane w postaci par atrybut-wartość, co wymaga rozszerzenia procesu przetwarzania żądania o dodatkowe kroki, którymi są co najmniej: dekodowanie (parsowanie), weryfikacja, obsługa błędów, interakcja ze źródłem danych (zasilanie danymi). W dalszej części odpowiedź na żądanie poprzedzona jest rozszerzonym w stosunku do pierwotnego modelu procesem generowania treści.

Inne różnice pojawiające się w omawianym modelu wbrew pozorom mają charakter następstw omawianych powyżej; same w sobie nie przesądzają jednak o tym, czy dana strona internetowa jest reprezentantem modelu. Do różnic takich zaliczyć można: występowanie formularzy w treści strony, przesyłanie żądań do serwera za pomocą metody POST[35], a nie właściwej dla zwykłych hiperłączy metody GET oraz utrudnioną indeksowalność treści [Kaczmarek2006].

2.1.3 Serwisy z zaawansowanym GUI

Model serwisów z zaawansowanym graficznym interfejsem użytkownika (GUI) charakteryzuje się rozszerzonymi funkcjonalnościami w zakresie interakcji z użytkownikiem w stosunku do modelu podstawowego. Historycznie rzecz ujmując, standard (X)HTML miał gwarantować tylko podstawowe wsparcie interakcji człowiek-komputer. Strony WWW miały na celu łatwe rozpowszechnianie informacji na masową skalę oraz umożliwienie nawigacji w przestrzeni dokumentów[36]. Rozwój rynku, w tym wzrost znaczenia firm z branży e-biznesu, spowodował jednak silną presję na ewolucję tego stanu rzeczy dla osiągnięcia dwóch celów. Po pierwsze: zwiększenia wygody użytkownika podczas korzystania z witryn WWW. Po drugie: zapewnienia jak najbardziej zbliżonego poziomu komfortu pracy użytkownika dla nowego modelu

[35] Nie jest to wymóg aczkolwiek ze względów praktycznych jest to najczęstsze rozwiązanie.
[36] http://www.w3.org/History/1989/proposal.html, odczytano 20-11-2012 r.

źródła internetowego – aplikacji webowych[37], w porównaniu do tradycyjnie uruchamianych lokalnie na komputerach aplikacji desktopowych.

Tradycyjnie pojmowany interfejs tworzony za pomocą stron WWW posiadał szereg wad, jeśli chodzi o osiąganie powyższych celów. Co do zasady, strony WWW były statyczne. Poza tym występowała konieczność przeładowywania całej strony nawet, jeśli zmiany w treści były niewielkie. Wreszcie brakowało standardowych elementów interfejsu użytkownika oraz mechanizmów z nimi związanych, do których przyzwyczajeni zostali użytkownicy aplikacji desktopowych (np. wsparcie techniki przeciągnij-i-upuść, rozwijane drzewa, okna, animowane obiekty etc.).

Dla przezwyciężenia wymienionych powyżej braków stworzono nowe standardy coraz lepiej wspierane przez kolejne wersje klientów WWW, co nie obyło się przy okazji bez sporych utrudnień i konfliktów[38]. Standardami tymi są: dynamiczny HTML umożliwiający osadzanie skryptów[39], asynchroniczna komunikacja z serwerem WWW, wsparcie dla dodatkowych standardów takich jak SVG[40]. Najpopularniejszym językiem skryptowym wykorzystywanym do rozszerzenia funkcjonalności po stronie klienta WWW jest język JavaScript[41].

Asynchroniczna komunikacja[42] z serwerem rozwiązuje istniejący przed jej pojawieniem się problem konieczności pobierania każdorazowo treści całej strony WWW w przypadku wprowadzania modyfikacji w tejże treści lub prezentacji nowej treści przy zachowaniu części treści poprzedniej. Istnieje szereg sposobów implementacji tego mechanizmu - obecnie najbardziej rozpowszechniony jest mechanizm programowej obsługi obiektu XMLHttpRequest, który w ramach implementacji przez każdą z przeglądarek internetowych umożliwia wykonywanie operacji wymiany danych z serwerem WWW[43]. Wymiana taka odbywa się za pomocą standardowego protokołu

[37] Aplikacji sieci Web, o których mowa będzie w dalszej części rozdziału.
[38] Chodzi przede wszystkim o zgodność z pojawiającymi się standardami oraz konkurencję pomiędzy producentami klientów WWW.
[39] Chodzi przede wszystkim o implementację standardu z serii ECMAScript, ale także np. rzadziej spotykane VBScript.
[40] ang. Scalable Vector Graphics, http://www.w3.org/TR/SVG/, odczytano 20-11-2012 r.
[41] Jest to w gruncie rzeczy nazwa najpopularniejszej implementacji wspomnianego już standardu ECMAScript.
[42] Mechanizmy komunikacji opisane w tym fragmencie są powszechnie określane akronimem AJAX (Asynchronous JavaScript and XML).
[43] Jednym z podstawowych powodów problemów związanych z tworzeniem dynamicznych stron oraz serwisów WWW a co za tym idzie także z ich analizą jest duże zróżnicowanie w producentów przeglądarek internetowych w przestrzeganiu standardów oraz indywidualne rozwiązania. Sytuację w tym zakresie poprawić mają nowe

HTTP(S), dając dowolność w zakresie formatów przesyłanych danych. Najczęściej są to: format tekstowy o określonej niestandardowej strukturze, dokumenty XML, inne formaty „lekkie" lub stworzony specjalnie w tym celu JSON[44]. Rzadziej stosuje się bardziej zaawansowane techniki kodowania przesyłanych danych – np. BASE64.

Serwisy z zaawansowanym GUI obecnie w większości buduje się przy użyciu gotowych szkieletów lub bibliotek[45].

2.1.4 Aplikacje webowe

Aplikacje webowe są modelem źródła internetowego charakteryzującym się realizacją skomplikowanych funkcjonalności wykraczających poza funkcjonalności zwyczajowo przypisane witrynom internetowym, czyli w szczególności nawigowaniu pomiędzy dokumentami. Zazwyczaj dla realizacji takich zaawansowanych funkcjonalności aplikacje webowe wykorzystują dodatkowe instrumenty dla ich właściwej implementacji. Jednym z takich instrumentów jest stanowość realizowana za pomocą mechanizmu sesji oraz za pomocą mechanizmów identyfikacji użytkownika. Do grupy takich instrumentów zaliczyć także można zaszycie w źródle internetowym określonej warstwy logiki biznesowej, która zapewnia nadzór nad wykonywanymi przez użytkowników akcjami, a także gwarantuje spójność realizacyjną pomiędzy poszczególnymi funkcjonalnościami. Względnie często modelowi omawianych źródeł internetowych towarzyszą elementy opisane w ramach modelu poprzedniego. Skomplikowane funkcjonalności aplikacji webowe wsparte są wówczas przez zaawansowane elementy interfejsu użytkownika. Połączenie takie daje lepsze wrażenie interakcji oferowane użytkownikowi aplikacji webowej, współgra ono także z możliwością wykorzystania gotowych ram i bibliotek oferujących rozwiązania stosowane zarówno po stronie klienta, jak i serwera. Możliwe jest jednak również zastosowanie takich ram wyłącznie po stronie serwera.

---

inicjatywy zmierzające do tworzenia bibliotek testów, takich jak: http://test262.ecmascript.org/ lub http://www.webstandards.org/action/, odczytano 20-11-2012 r.

[44] JavaScript Object Notation, http://www.json.org/, odczytano 20-11-2012 r.

[45] Listę i porównanie popularnych rozwiązań tego typu można znaleźć w: http://en.wikipedia.org/wiki/List_of_web_application_frameworks oraz http://en.wikipedia.org/wiki/List_of_Ajax_frameworks, odczytano 20-11-2012 r.

2.1.5 Serwisy spersonalizowane

Personalizacja treści w witrynach WWW jest pomysłem, który pojawił się stosunkowo wcześnie przy projektowaniu portali internetowych. W przypadku serwisów zasilanych danymi oznacza ona sytuację, w której prezentowane treści są funkcją użytkownika żądającego dostępu do witryny lub pewnych cech związanych z tym użytkownikiem. Zakłada się, że dla serwisów niepodlegających personalizacji, dostarczana treść jest niezależna od użytkownika oraz jego wykrywalnych cech.

Serwisy spersonalizowane polegają na pojęciu profilu użytkownika. Profil taki jest tworzony w oparciu o informacje dostarczone przez samego użytkownika (preferencje, dane demograficzne etc.) wraz z odpowiednimi mechanizmami uwierzytelnienia. Informacje do profilu mogą być także gromadzone w oparciu o zachowania użytkownika (profilowanie behawioralne) lub mogą być dostarczane automatycznie w ramach infrastruktury sieci WWW. Do mechanizmów automatycznych zaliczyć można: wymianę informacji przez klienta WWW, geolokalizację na podstawie adresów IP, ew. dostęp do informacji zapisanej w plikach z ciasteczkami[46] lub historii nawigacji pod warunkiem nienaruszania prywatności.

2.1.6 Serwisy e-commerce

Jest to model źródeł internetowych bezpośrednio uczestniczących w obrocie gospodarczym. Źródła takie mają za zadanie prezentowanie oferty handlowej jednego lub wielu podmiotów gospodarczych w zakresie sprzedaży dóbr lub usług. Poza samą ofertą dostarczają one także informacji w zakresie warunków transakcji kupna lub sprzedaży. Wreszcie pozwalają także zawrzeć same transakcje.

2.1.7 Pozostałe modele źródeł internetowych

Do innych modeli serwisów zasilanych danymi zaliczyć należy: serwisy adaptatywne, serwisy wykorzystujące filtrowanie grupowe, portale bankowe[47], usługi webo-

[46] ang. cookies.

[47] Jednym z najstarszych i zarazem najbardziej popularnym serwisem e-banking w Polsce jest witryna https://www.mbank.com.pl/, mBank jest także jednym z największych pośredników ubezpieczeniowych za pomocą kanału bankowego (bankassurance) on-line. Wg danych Gemius Megapanel za sierpień 2012 r. cała witryna miała 1 875 391 użytkowników.

we[48]. Modele te zasadniczo nie stanowią bezpośredniego przedmiotu zainteresowania z punktu widzenia niniejszej pracy, ich cechy bowiem stosunkowo rzadko pojawiają się w przypadku witryn mogących stanowić źródła wiedzy ubezpieczeniowej tak, jak zostało to zdefiniowane wcześniej.

2.2 Ekstrakcja informacji ze źródeł internetowych

2.2.1 Pojęcie ekstrakcji informacji

Ekstrakcję informacji (EI) ze źródeł webowych definiuje się jako „zautomatyzowaną transformację stron WWW do postaci ustrukturyzowanych danych" [Chang2006]. Przytoczona definicja[49] opiera się na założeniu, że proces ekstrakcji na wejściu pobiera informację bez struktury lub słabo ustrukturyzowaną, natomiast na wyjściu zwraca postać z pełną i ustaloną strukturą. Nie ma, co prawda, zgody co do tego, czym jest „postać ustrukturyzowana", niemniej wielu autorów (np. [McCallum2002]) utożsamia ją ze strukturami analogicznymi do formy relacji znanej z baz danych [Codd1970]. Jeżeli dodatkowo proces ekstrakcji obejmuje uzgadnianie informacji pochodzących ze źródeł o różnorodnych reprezentacjach i ujednoznacznienie w związku ze sprowadzeniem ich do wspólnej postaci, to można mówić o zadaniu integracji informacji ze źródeł webowych[50] [Kaczmarek2006].

Typologia[51] procesu ekstrakcji informacji obejmuje wykorzystanie zróżnicowanych kryteriów. Przykładowo [Hsu1998] klasyfikują systemy EI jako: ręcznie stworzone osłony[52] z użyciem powszechnie dostępnych języków programowania, ręcznie stworzone osłony ze specjalistycznymi językami, osłony wykorzystujące heurystyki oraz osłony indukcyjne. [Kuhlins2002] wprowadza prozaiczny podział na rozwiązania

[48] ang. web service. Termin tłumaczy się także jako "usługa internetowa", http://www.w3.org/TR/ws-gloss/, odczytano 20-11-2012 r.
[49] Wskazać można na kilka innych definicji znajdujących się w literaturze. Są one jednak równoważne.
[50] Jest to zatem proces rozszerzony w stosunku do oryginalnego zakresu ekstrakcji informacji. Zadanie integracji informacji wykracza jednak poza zakres pracy.
[51] Bardzo obszerny opis prac związanych z ekstrakcją informacji ze źródeł webowych znaleźć można w dysertacji [Flejter2011]. Znajduje się tam m.in. rozbudowana klasyfikacja tego rodzaju systemów.
[52] ang. wrapper – osłona. Chodzi o komponenty programowe zapewniające określony poziom abstrakcji i pośredniczące pomiędzy źródłem informacji a mechanizmem obsługi zapytań. Pół- lub całkowicie automatyczne ekstrahowanie informacji ze źródeł sieci Web wymaga wewnętrznego mechanizmu reprezentacji takiego źródła. Osłony pozwalają na uogólnione podejście do takiej reprezentacji. Szerzej będzie o nich mowa w kolejnym podrozdziale.

komercyjne oraz niekomercyjne[53]. Z kolei [Kushmerick2003] wyróżnia podział na: systemy skończeniestanowe i stosujące uczenie relacyjne. Propozycja [Chang2006] stanowi syntezę różnych systemów klasyfikacji rozwiązań służących do realizacji zadań ekstrakcji informacji z internetu.

W gruncie rzeczy metody ekstrakcji stosowane w przypadku poszczególnych rodzajów źródeł sprowadzają się ostatecznie do umiejętnego rozpoznawania wzorców w ciągach danych. W przypadku informacji o bardziej regularnej strukturze mniej wyrafinowane sposoby rozpoznawania wzorców przynoszą zadowalające rezultaty, jednocześnie zdecydowanie zmniejszając koszt zużycia zasobów podczas realizacji procesu rozpoznawania.

Dostatecznie zaawansowane systemy ekstrakcji informacji posiadają cechy zarówno rozwiązań ekstrakcji webowej, jak i tekstowej[54]. Dobrym przykładem mogą być tutaj rezultaty projektów [Węcel2011] oraz [Kaczmarek2010]. W architekturach obu tych projektów poza metodami ekstrakcji strukturalnej informacji, charakterystycznej dla ekstrakcji ze źródeł internetowych, zastosowano także moduły ekstrakcji leksykalnej oparte na, co prawda względnie prostych, ale jednak mechanizmach charakterystycznych dla zastosowań znanych z dziedziny przetwarzania języka naturalnego[55]. Podejście takie spowodowało wielokrotne zwiększenie efektywności[56] i jakości rezultatów działania tychże systemów. Z drugiej strony oczywiście podyktowane było specyficznymi potrzebami polegającymi na przetwarzaniu określonych typów dokumentów[57], ale podkreślmy – dokumentów ekstrahowanych z sieci WWW.

2.2.2 Najważniejsze systemy ekstrakcji informacji ze źródeł internetowych

Jak wynika z przytoczonej na początku poprzedniego rozdziału definicji, ekstrakcja informacji jest w istocie procesem nadawania lub formalnego wyspecyfikowania

[53] Większość opisywanych w pracy projektów ma charakter badawczy i niekomercyjny. Przykładami aplikacji komercyjnych są: Jango [Doorenbos1997] – będący w założeniu twórców dziedzinowo niezależnym pośrednikiem zakupowym oraz system iMacros [Iopus2012] – służący do tworzenia makr automatyzujących nawigację za pośrednictwem przeglądarki internetowej.
[54] Chodzi tu o rozróżnienie procesu ekstrakcji ze źródeł częściowo ustrukturyzowanych (dokumenty HTTP) oraz praktycznie nie posiadających struktury (tekst).
[55] Konkretnie zastosowano obróbkę tekstu za pomocą zbiorów gramatyk zapisanych w formalizmie JAPE stworzonym na potrzeby środowiska przetwarzania tekstu GATE.
[56] Efektywność rozumiana jest tutaj nie jako czas przetwarzania, ale stopień wykrywalności określonych wzorców.
[57] Chodzi o wpisy na portalach o charakterze społecznościowym lub prezentacje personalne. Zasadniczo jednak dokumenty tworzone przez ludzi, nie zaś generowane przez automaty.

struktury. Jeśli chodzi o podejścia do strukturyzacji informacji w źródłach webowych, to wyodrębnić można dwa zasadnicze nurty [Iskold2007]:

- oddolny – twórcy treści (stron) są odpowiedzialni za oznaczenie tekstu tak, żeby był łatwo przetwarzalny w sposób automatyczny,
- odgórny – sposoby publikowania informacji w sieci pozostają niezmienione, natomiast powszechnie stosowane są algorytmy, których celem jest wykrywanie i akwizycja informacji.

W przypadku podejścia oddolnego wykorzystanie znajdują wszelkiego rodzaju formalizmy, takie jak: XML, XSLT [Clarck1999], RSS [Rss2007], RDF [Beckett2004], RDFa, OWL [McGuinness2004], JSON, DublinCore[58], mikroformaty[59] i inne techniki wspomagające strukturyzację informacji. Przy czym zauważyć trzeba, że użycie tych formalizmów, z wyjątkiem dość powszechnych mikroformatów[60], jest względnie rzadkie. Powodów takiego stanu rzeczy jest wiele. Przede wszystkim twórcy traktują człowieka jako podstawowego odbiorcę treści. Dodatkowo duża część treści w sieci powstała w czasie, kiedy nie stawiano sobie ambitnych celów związanych z automatyzacją przetwarzania. Nie bez znaczenia jest również fakt, że wiele organizacji oraz podmiotów komercyjnych (m.in. w sektorze handlu elektronicznego) dąży do utrudnienia dostępu do informacji odbiorcom niebędącym bezpośrednimi klientami lub zmierzających do wykorzystania informacji w sposób niezgodny z zaplanowanym modelem biznesowym[61].

Nurt odgórny nakłada wymóg powstania specjalistycznych systemów i oprogramowania realizującego zadanie ekstrakcji informacji. Historycznie jednym z pierwszych systemów, które uznać można za wczesną wersję rozwiązania do ekstrakcji informacji webowej, był TSIMMIS[62] [Chawathe1994]. Pierwsze doniesienia o tym projekcie pojawiają się w 1994 roku, czyli niespełna 4 lata po powstaniu sieci WWW[63]. Ten kilkuletni projekt realizowany wspólnie przez Uniwersytet Stanforda oraz IBM

[58] http://dublincore.org, odczytano 20-09-2013 r.
[59] http://microformats.org, odczytano 20-09-2013 r.
[60] vCard,
[61] Tym samym dążąc do utrzymania status quo w zakresie wspominanej na początku pracy asymetrii informacji.
[62] The Stanford-IBM Manager of Multiple Information Sources.
[63] Pierwszy serwer HTTP rozpoczął pracę w CERN w grudniu 1990 r. [Berners-Lee2000]

Almaden Research reprezentuje w istocie bardziej cechy ogólnej ramy[64] do integracji informacji z heterogenicznych źródeł niż narzędzie przeznaczone wyłącznie do procesu akwizycji. Prace prowadzone w ramach TSIMMIS wprowadziły jednak koncepcję ogólnego języka zapytań zbliżonego do SQL, w którym źródłem rezultatów mogła być strona WWW oraz formalizm zapisu manualnie tworzonych reguł transformacji HTML do struktur obiektowych.

Mechanizm ręcznego tworzenia reguł ekstrakcji jest właśnie cechą wielu wczesnych systemów ekstrakcji, które częściowo powielały schemat powyżej omówionego rozwiązania. Do tej kategorii systemów zaliczyć można także projekt WebSQL [Mendelzon1997]. Podobny język zapytań wprowadzał projekt W3QS [Konopnicki1995], z tym że jednocześnie rozszerzał on możliwości nawigacyjne po źródle o zautomatyzowane wypełnianie formularzy. Inną grupę języków zapytań, tym razem nawiązujących do różnych formalizmów logicznych, stanowią takie rozwiązania jak: WebLog [Florescu1997] czy FLORID [Himmeröder1997]. Ten ostatni wprowadzał aparat pojęciowy bazujący na logice ram (F-logic). Z kolei istotnym uogólnieniem większości podejść występujących w tym nurcie była praca Web-OQL [Arocena1998].

Przykładem zgoła innego podejścia do problemu ekstrakcji na dużą skalę jest WIEN[65] [Kushmerick1997]. W odróżnieniu od systemów czysto manualnych wspomniane rozwiązanie stworzone było jako prototyp architektury w pełni zautomatyzowanej. Środowisko systemu pozwalało na zastosowanie szeregu osłon – od bardzo prostych do średnio skomplikowanych mechanizmów rozpoznających wzorce wynikające ze współwystępowania znaczników. Za następców systemu WIEN uznać można takie projekty jak: SoftMealy [Hsu1998], który jako jeden z pierwszych systemów wykorzystywał specyficzne typy automatów skończonych, XWRAP [Liu2000], czy Stalker [Muslea1999], który operował na poziomie tokenów oraz znaczników.

Trzecią istotną grupą systemów ekstrakcji informacji ze źródeł webowych stanowią silniki opakowujące, działające na danych dostarczonych przez użytkownika przez specjalny interfejs. Pierwszym rozwiązaniem, które realizowało taką ideę, był NoDoSE [Adelberg1998]. Dostatecznie ogólna architektura pozwalała w nim na ekstrakcję

[64] ang. framework
[65] Wrapper Induction ENvironment

informacji zarówno z tekstu, jak i z dokumentów HTML[66]. Rozwiązanie z założenia stanowić miało ramę do testowania różnych algorytmów budowy osłon. Struktura dokumentu jako takiego reprezentowana była w postaci specyficznego drzewa. Zadaniem użytkownika było wskazanie fragmentów dokumentu. Na tej podstawie generowane były reguły wykrywania podobieństwa pomiędzy strukturą „leżącą pod" treściami wskazanymi przez użytkownika oraz pozostałą strukturą dokumentu. Następnie reguły te były ulepszane. Do tej samej grupy systemów zaliczyć można m.in. prace dotyczące systemu W4F[67] [Azavant1999] i rezultaty opisane w pracy [Ashish1997]. Nowatorstwem we wspomnianym projekcie W4F było zastosowanie języka HEL[68], przypominające uproszczoną wersję obecnie szeroko rozpowszechnionego języka XPath[69]. Bardziej zaawansowana idea rozwiniętego interfejsu użytkownika przeznaczonego do wizualnej budowy osłon zaprezentowana została w ramach prac prowadzonych na Federalnym Uniwersytecie Minas Gerais[70]. Nieco późniejszym rozwiązaniem, ale działającym na podobnych założeniach, był Lixto [Baumgartner2001] – system rozwijany na Politechnice Wiedeńskiej, który dodatkowo wspierał nawigację w głębokim internecie[71].

Ostatnią silnie wyróżniającą się grupą systemów ekstrakcji informacji są osłony budowane za pomocą technik uczenia nienadzorowanego. Wzorcowym przykładem takiego podejścia jest projekt Exalg [Arasu2005].

Pracując nad algorytmem Exalg, autorzy zdecydowali się uwzględnić pewne założenia. Przede wszystkim ustalono model, w którym dokumenty generowane są przez źródło za pomocą gotowego szablonu. W konsekwencji szablon ten jest wypełniany danymi o określonym schemacie. Schemat danych składa się wyłącznie z typów prostych[72] oraz dwóch rodzajów struktur: list i zbiorów. Struktury identyfikowane są za pomocą konstruktorów – specyficznych operatorów działających na typach pro-

[66] To ostatnie jest jednak dość ułomne.
[67] World-Wide Web Wrapper Factory
[68] HTML Extraction Language.
[69] http://www.w3.org/TR/xpath/, odczytano 20-11-2012 r.
[70] Mowa o projekcie Data Extraction By Examples (DEByE) oraz jego dalszych kontynuacjach.
[71] Opis samego źródła w tym systemie zrealizowany został z wykorzystaniem autorskiego języka Elog wywodzącego się z logicznego formalizmu Datalog, co zbliża to rozwiązanie do pierwszej z wyróżnionych grup systemów ekstrakcji informacji. Dodatkowo nawigacja wykorzystywała także język XPath, co było pewnym nowatorstwem.
[72] Odpowiadających w zasadzie wyłącznie ciągom znaków w postaci słów lub znaczników.

stych lub podstrukturach. W ramach modelu szablonu wyróżnić można z kolei funkcję, która rekurencyjnie realizuje zamieszczenie danych ze schematu w gotowych dokumentach. Funkcjonowanie samego algorytmu odbywa się w dwóch etapach. Pierwszy etap związany jest z wyznaczaniem klas równoważności. Autorzy wprowadzają to pojęcie, definiując klasę równoważności jako zbiór tokenów[73] występujących z taką samą częstością w każdym dokumencie pochodzącym z analizowanego źródła. Ostatecznie tylko szczególne przypadki[74] wyznaczonych klas równoważności są przedmiotem dalszego przetwarzania przez algorytm. Rozpatrywane w pierwszym kroku klasy są dodatkowo różnicowane ze względu na umiejscowienie w strukturze dokumentu oraz kontekst poszczególnych tokenów składających się na dane klasy. Drugi etap działania algorytmu polega na generowaniu szablonu, wykorzystując informacje o zidentyfikowanych klasach równoważności.

Do omówionej jako ostatnia kategorii systemów EI zaliczyć można również system RoadRunner [Merialdo2001], będący produktem współpracy dwóch włoskich uczelni – Uniwersytetu w Rzymie oraz Universita della Basilicata. W tym rozwiązaniu formalizmem opisu wzorców były wyrażenia regularne.

2.2.3 Wyzwania dla systemów ekstrakcji informacji a odkrywanie wiedzy ubezpieczeniowej ze źródeł internetowych

W przypadku skomplikowanych źródeł sieci Web reprezentacja źródła wymaga bardziej wyrafinowanego podejścia. Sytuacja taka dotyczy przede wszystkim źródeł głębokiego internetu oraz serwisów o zaawansowanym interfejsie użytkownika (GUI). Źródła głębokiego internetu generują dodatkowo problem nawigacji przez formularze [Shestakov2005, Kaczmarek2006]. Z kolei nawigowanie po źródłach z zaawansowanym GUI wymaga m.in. pokonania wyzwania, jakim są dynamicznie zmieniane treści [Alvarez2007].

Wyzwań, które muszą być pokonane w celu przeprowadzenia pełnego i zakończonego sukcesem procesu ekstrakcji danych, jest oczywiście znacznie więcej. Znaczna

[73] ang. symbol – w przypadku przetwarzania języka naturalnego pojęciem tym często określa się podstawowe jednostki analizowanego dokumentu.
[74] Chodzi o klasy dostatecznie duże pod względem liczebności tokenów oraz dostatecznie często występujące w zbiorze dokumentów źródłowych.

część z nich ma charakter szczegółowych rozwiązań technicznych[75]. Bardzo wyczerpujące i usystematyzowane wyliczenie wyzwań zawarte zostało w przytaczanej już pracy [Flejter2011]. Autor przedstawia w niej 336 problemów, z którymi musi zmierzyć się nowoczesny system przy realizacji zadania ekstrakcji informacji ze źródła webowego.

Tabela 1. Rozszerzona lista wyzwań dla nowoczesnych systemów ekstrakcji informacji
Źródło: opracowanie własne na podstawie [Flejter2011]

III		
III.1	Efekt „patrzenia przez dziurkę od klucza”	
Możliwości zapytania – wiele atrybutów		
III.1.a	Zmienne zakresy atrybutów, zmienne kategorie atrybutów	+
III.1.b	Obsługa zależności pomiędzy atrybutami	+
IV	Wyzwania związane z logiką aplikacji webowych po stronie klienta	
IV.2	Rozpoznanie (interpretacja) elementów interfejsu użytkownika	
Parsowanie formularzy		
IV.2.a	Warunkowy dostęp do części formularza	+
IV.2.b	Schemat formularzy zależny od wartości atrybutów	+
IV.2.c	Modyfikowalne formularze (zależne od interakcji użytkownika)	+
VI	Ekstrakcja danych	
VI.2	Ekstrakcja danych złożonych	
VI.2.a	Łączenie wartości atrybutów z ekstrahowanymi danymi	+
VII	Planowanie nawigacji	
VII.3	Wybór właściwego zakresu nawigacji	
Strategie pracy ze źródłami zawierającymi zbyt dużo danych		
VII.3.a	Ustalenie zależności pomiędzy atrybutami a ekstrahowanymi danymi	+
VII.3.b	Dobór wartości dla atrybutów	+
VII.3.c	Wykorzystanie wiedzy zewnętrznej dla optymalizacji nawigacji	+

Z powodu znacznego stopnia skomplikowania i konstrukcji większości portali ubezpieczycieli[76] przy realizacji zadania wyzwania te są aktualne również w przypadku proponowanego zadania i metody jego rozwiązania. Przywołana lista problemów podzielona jest na 2 poziomy kategorii. Najwyższy poziom podziału złożony jest z 9 pozycji: różne modele danych, interakcja z serwerami, logika aplikacji po stronie ser-

[75] Nie oznacza to, że można je mimo to lekceważyć, ponieważ nawet błahe przeszkody mogą w praktyce uniemożliwić realizację omawianego zadania.
[76] Podstawowy podział technologii został zaprezentowany w podrozdziale 2.1. Natomiast o ich faktycznym wykorzystaniu w ramach portali dostarczających wiedzę ubezpieczeniową mowa będzie w rozdziale 4.

wera, logika aplikacji po stronie klienta, różne typy treści, ekstrakcja danych, planowanie nawigacji, wykonanie nawigacji, specyfika scenariuszy użycia.

Ze względu na specyfikę projektowanego przez nas zadania proponujemy rozszerzenie przywołanej hierarchii wyzwań o dodatkowe elementy. Propozycja, stanowiąca uzupełnienie listy podanej w [Flejter2011], zawarta została w tabeli 1. Dodane elementy oznaczono symbolem „+".

2.3 Ekstrakcja wiedzy i metody eksploracji danych

Tematyką znajdującą się na pograniczu ekstrakcji informacji ze źródeł internetowych oraz eksploracji danych są trzy pokrewne zagadnienia określane zbiorczo pojęciem Web mining[77]. Do zagadnień kryjących się pod tym terminem zaliczają się: eksploracja struktury witryn, eksploracja danych z treści witryn, eksploracja informacji o użyciu witryn[78].

Eksploracja danych webowych definiuje się jako „zastosowanie technik odkrywania wiedzy do automatycznego odkrywania i ekstrahowania informacji z dokumentów i usług webowych" [Kosala2000]. Zaznaczyć należy, że przytoczona definicja jest poprawna w stosunku do trzeciego i częściowo pierwszego z omawianych w ramach tego podrozdziału zagadnień. Dzieje się tak dlatego, że zagadnienie drugie, a częściowo także i pierwsze, porusza raczej kwestie operowania na metadanych – rozumianych tutaj jako informacje o witrynie, niekoniecznie lub czasem na pewno nie znajdujące się bezpośrednio w ramach treści dokumentów na niej zamieszczonych. Oznacza to, że takie metadane pochodzić będą z innego źródła niż sama analizowana witryna[79]. W pracy wykorzystamy rozwiązania nawiązujące do eksploracji danych webowych.

Eksploracja struktury witryn

Eksploracja struktury witryn polega na analizie i wyciąganiu użytecznych wniosków dotyczących struktury powiązań pomiędzy dokumentami zamieszczonymi wewnątrz danej witryny oraz dokumentami zewnętrznymi[80]. Dokumenty w sieci WWW są węzłami w grafach, a wiedza o ich cechach oraz topologii stanowi często cenny zasób biznesowy oraz interesujący przedmiot badawczy. W szczególności wiedza

[77] Eksploracji danych webowych.
[78] ang. odpowiedniki kolejno to: Web structure mining, Web content mining, Web usage mining.
[79] np., logi z serwera w przypadku analizy danych o użyciu witryn.
[80] Znajdującymi się w innych witrynach lub domenach.

ta może sprowadzać się do wskazówek odnośnie ważności dokumentów lub całej witryny w sieci – dobrymi przykładami zastosowania są tutaj algorytmy wspierające systemy wyszukiwawcze, takie jak PageRank [Page1998] wraz z jego rozwinięciami czy wcześniejszy HITS[81].

Istotnym zastosowaniem analizy struktury jest również mierzenie stopnia skomplikowania witryn[82] oraz - rzadziej spotykana – kategoryzacja witryn w oparciu o strukturę. Inne zastosowania obejmują także zagadnienie detekcji zgrupowań oraz społeczności tematycznych, czemu poświęcone było również szereg badań.

Eksploracja danych z treści witryn

Stanowi kategorię badań w zakresie Web mining najsilniej skojarzoną z zaprezentowaną w kolejnych rozdziałach metodą ekstrakcji modeli wyceny składki. Na typowe zadania zawierające się w tym nurcie składa się sekwencja działań polegająca na identyfikacji zasobu (dokumentu), selekcji i przetworzenia informacji, wykryciu występujących wzorców poprzez analizę treści oraz sprawdzeniu poprawności i interpretacji zagregowanej wiedzy.

Eksploracja danych z treści witryn ma szereg praktycznych zastosowań, chociaż względnie rzadko metody takiego odkrywania wiedzy występują samodzielnie. Znacznie częściej są one wykorzystywane na potrzeby realizacji celów wytyczanych w ramach innych pokrewnych dziedzin, takich jak: ekstrakcja informacji czy wyszukiwanie informacji. Taki stan rzeczy, jak również trudności w ostrym wyznaczeniu zakresu metod drążenia danych[83], powoduje problemy w jednoznacznej klasyfikacji prac prowadzonych w ramach opisywanego nurtu. W jednym z opracowań podjęto próbę podsumowania prowadzonych prac przez wprowadzenie podziału poszczególnych podejść na perspektywy: wyszukiwania danych nieustrukturyzowanych, ustrukturyzowanych oraz bazodanowej. Zdecydowanie prostszym sposobem podsumowania prac wydaje się jednak kryterium realizowanych zadań.

[81] Hyperlink-Induced Topic Search [Kleinberg1999]

[82] Zagadnienie istotne praktycznie ze względu na konsekwencje marketingowe w przypadku witryn o charakterze komercyjnym.

[83] Przykładowo autorzy [Kosala2000] rozróżniają metody data mining-u oraz uczenia maszynowego choć wydaje się, że niektóre metody uczenia maszynowego należą do standardowego arsenału podejść w drążeniu danych – np. sztuczne sieci neuronowe.

W ramach prac prowadzonych w nurcie eksploracji danych z witryn wyszczególnić można następujące zadania [Zhang2008]:

- grupowanie, kategoryzacja lub klasyfikacja treści,
- identyfikacja i wzbogacanie treści[84],
- generowanie skrótów i podsumowań,
- odkrywanie pojęć, relacji, budowa słowników oraz ontologii,
- zastosowania społeczne – odkrywanie opinii, emocji etc.

Wspomniany wcześniej podział na treści ustrukturyzowane oraz nie posiadające struktury jest o tyle istotny, że powyżej wynotowane zadania realizuje się z użyciem nieco odmiennych metod dla obydwu typów informacji. Nowo powstającym, ale posiadającym spory potencjał, jest nurt związany z jeszcze innym rodzajem danych, a mianowicie z danymi multimedialnymi[85].

Zadanie i metoda jego realizacji stanowiąca przedmiot niniejszej pracy wpisuje się w nurt badań prowadzonych w ramach eksploracji danych z treści witryn. Zwłaszcza dobrze dopasowana jest do definicji omawianego podejścia ekstrakcji wiedzy[86]. Mimo to sformułowanie problemu badawczego w takiej wersji nie nastąpiło wcześniej w ramach omawianego nurtu, a co za tym idzie, nie ma doniesień o próbach realizacji podobnego zadania.

Eksploracja informacji o użyciu witryn

Najbardziej odmiennym spośród trzech omawianych podejść eksploracji danych webowych jest eksploracja informacji o użyciu witryn. Idea tego podejścia sprowadza się do wykorzystania różnego rodzaju informacji, w szczególności nie zawartej bezpośrednio w treści witryny, do odkrycia zależności i wzorców związanych z użytkowymi aspektami korzystania z witryny. Wzorce używania i nawigowania po lub pomiędzy witrynami mają istotne znaczenie w wielu zastosowaniach. Warte zauważenia jest, iż wzorce takie wykorzystywane są również w zadaniach ekstrakcji informacji z internetu oraz w zadaniu, którego metoda realizacji stanowi przedmiot prezentowanej pracy.

[84] Poprzez automatyczne adnotowanie dokumentów.
[85] Obrazy, strumienie dźwiękowe oraz filmy.
[86] „Drążenie danych z witryn opisuje odkrywanie użytecznej informacji z treści webowej/danych/dokumentów. Niemniej, to co składa się na treści webowe stanowi bardzo szeroką gamę danych„ [Kosala2000].

2.4 Eksploracja danych

Metody eksploracji danych (data mining) służą do odkrywania ukrytych wzorców, powiązań i trendów przez przeszukiwanie dużych ilości danych z wykorzystaniem odpowiednich metod statystycznych.

Wyróżniamy dwa główne podejścia do data mining: uczenie kontrolowane i uczenie niekontrolowane. To drugie obejmuje szukanie zależności w samych danych. Do celów podejmowania decyzji dużo ważniejsze jest jednak pierwsze podejście, które pozwala na zbudowanie modelu decyzyjnego.

Można wskazać cztery podstawowe problemy rozwiązywane z wykorzystaniem metod data mining [Berry2000]. Pierwszy z nich to dokonywanie wyborów, czyli klasyfikacje, dla których szczególne zastosowanie mają drzewa decyzyjne [Quinlan1986]. Drugi problem to tworzenie prognoz, czyli przybliżanie nieznanej przyszłej wartości. Tutaj rozwiązania opierają się na budowie numerycznych modeli predykcyjnych przybliżających postać określonych funkcji. Użyteczne są tu m.in. metody uczenia sztucznych sieci neuronowych, metody ewolucyjne, czy też analiza regresji. Z punktu widzenia problemu naukowego niniejszej pracy zagadnienie tworzenia takich modeli predykcyjnych jest szczególnie istotne. Trzeci problem to odkrywanie relacji w danych (np., analiza skupień [Bramer2007]), a czwarty to usprawnianie procesów[87].

2.4.1 Regresja

Regresja jest metodą statystyczną służącą do tworzenia matematycznych modeli opisujących zbiory danych przez określenie związku pomiędzy ich wyodrębnionymi podzbiorami. Odkrycie tych związków umożliwia predykcję wartości jednych wielkości na podstawie znanych wartości pozostałych wielkości.

W przypadku stosowania regresji parametrycznej[88] zakłada się posiadanie pewnej wiedzy a priori o procesie generującym dane. Posiadając taką wiedzę, możliwe jest stosowanie tzw. modeli parametrycznych, w których postać modelu jest dana z góry,

[87] Te zagadnienia nie są istotne dla dalszych rozważań.
[88] O szczególnych przypadkach regresji nieparametrycznej mowa będzie w kolejnym podrozdziale.

natomiast ustaleniu – estymacji – podlegają same parametry. Postać takiego modelu przedstawia się następująco:

$$Y = f(X, B) + \epsilon. \quad (1)$$

Celem zastosowania metody jest jak najlepsze przybliżenie nieznanej funkcji f za pomocą estymatora f'. W praktyce odbywa się to przez dostosowanie wektora współczynników B tak, aby dla zbioru uczącego funkcja błędu przyjmowała wartość minimalną. Najczęściej stosowaną i zarazem jedną z prostszych w zastosowaniu funkcji błędu jest miara sumy kwadratów różnic między wartościami teoretycznych i empirycznymi[89].

Do najczęściej stosowanych modeli parametrycznych zaliczyć należy: regresję liniową, regresję nieliniową oraz uogólnione modele liniowe (GLM).

2.4.2 Programowanie genetyczne

Na wysokim poziomie abstrakcji algorytmy ewolucyjne określić można jako symulację operującą na zbiorze bytów określanym jako **populacja**. Każdy byt – **osobnik** znajdujący się w przetwarzanym zbiorze – populacji, stanowi możliwy rezultat dla rozwiązywanego problemu. Problem ten określa konstrukcję i specyfikę środowiska, w którym osobniki funkcjonują oraz sposób opisu (kodowania) bytów. Ze względu na to środowisko istnieje charakterystyczna funkcja przyporządkowująca każdemu osobnikowi wartość liczbową. Wartość ta odzwierciedla jakość rozwiązania skojarzonego z danym osobnikiem i nazywamy ją **przystosowaniem osobnika**. Natomiast funkcja, która tę wartość przypisuje, stanowi **funkcję przystosowania**[90].

Kluczowymi pojęciami zaczerpniętymi przez analogię z nauk biologicznych w przypadku algorytmów ewolucyjnych są pojęcia genotypu oraz fenotypu[91]. **Genotypem** jest informacja skojarzona z każdym pojedynczym osobnikiem, stanowiąca przedmiot przetwarzania przez czynniki dostosowawcze w algorytmie. Postać genotypu determinuje **fenotyp**, czyli zespół cech danego osobnika. Cechy te wpływają bez-

[89] Ma ona tę wadę, że jest czuła na elementy danych istotnie odstających.
[90] Z punktu widzenia symulacji „środowisko" często redukowane jest praktycznie wyłącznie do właściwie zdefiniowanej funkcji przystosowania. Funkcja taka może mieć dowolną postać, dodatkowo może ulegać zmianom w trakcie działania algorytmu lub zawierać czynniki losowe.
[91] W uproszczonych symulacjach w szczególnych przypadkach wprowadza się równoważność pomiędzy genotypem osobników a ich fenotypami. Jest to kwestia kodowania tych informacji w określonej implementacji.

pośrednio na ocenę osobnika z punktu widzenia środowiska w postaci wartości funkcji przystosowania.

Informacja zapisana w genotypach osobników ma charakter ustrukturyzowany. Składają się na nią chromosomy oraz geny, przy czym **chromosom** jest zbiorem genów natomiast cały genotyp jest zbiorem chromosomów. **Geny** są traktowane jako niepodzielna jednostka dziedziczonej informacji. Dziedziczenie to odbywa się w cyklach. Każdy cykl polega na przekształceniu populacji zastanej w populację potomną tak, że populacja potomna staje się jednocześnie populacją zastaną w kolejnym cyklu wykonywania algorytmu. Algorytm ewolucyjny nie ulega zatrzymaniu, chyba że zostanie osiągnięty subiektywnie zdefiniowany warunek stopu.

Generowanie populacji potomnej odbywa się przez wybór osobników z populacji zastanej oraz zastosowanie na ich genotypach określonych operatorów genetycznych. Operatorami tymi klasycznie są: **mutacja** i **krzyżowanie**[92].

Algorytmy ewolucyjne podzielić można na algorytmy genetyczne [Arabas2001], strategie i programowanie ewolucyjne oraz programowanie genetyczne. Wszystkie one zakładają wykorzystanie mechanizmów symulowanego doboru naturalnego i dziedziczenia dla tworzenia potencjalnej przestrzeni rozwiązań. Pionierem klasycznych zastosowań dla algorytmów genetycznych jest A. Fraser [Fraser1957]. Koncepcja programowania genetycznego rozwinięta została przez J. Kozę [Koza1994]. Obecnie częstym zabiegiem jest wykorzystanie dla reprezentacji programów pseudokodu maszynowego (metoda AIM [Nordin1999])[93].

2.4.3 Sztuczne sieci neuronowe

Sztuczne sieci neuronowe są strukturami, które przekształcają sygnały wejściowe na sygnały wyjściowe przez mechanizm symulowanego przesyłania i modyfikowania sygnałów pomiędzy numerycznymi odpowiednikami komórek nerwowych. Podstawową cechą charakterystyczną omawianych struktur jest zdolność uczenia i rozpoznawania wzorców. Zasadniczo sieci neuronowe różnią się pomiędzy sobą: rodzajami i poziomem szczegółowości symulowanych neuronów [Osowski2000], liczbą neuro-

[92] ang. cross-over, [Ahmed2010]

[93] Może to dotyczyć zarówno natywnych języków maszynowych procesorów, ale również języków, w których przechowywany jest kod dla maszyn wirtualnych (Java, CLI). Oprogramowanie, wykorzystane do badań opisanych w rozdziale 7 korzysta z takiego rozwiązania.

nów, konfiguracjami połączeń pomiędzy neuronami [Miller1989] oraz metodami uczenia [Montana1989]. Różnice te są szczególnie istotne ze względu na konkretne zastosowania poszczególnych implementacji sieci.

Pierwsze matematyczne modele neuronów powstały w latach 40. [McCulloch1943]. Modele te od tego czasu są systematycznie ulepszane, uwzględniając specyficzne cechy funkcjonowania komórek nerwowych. W 1958 roku powstał perceptron [Rosenblatt1958], który był pierwszą symulowaną na komputerze siecią neuronową. Jednak dopiero w 1986 roku pojawił się przełom w badaniach w tym obszarze za sprawą pracy [Rumelhart1986]. Rozwinięto w niej metodę uczenia z propagacją błędu, co było istotnym krokiem na drodze do stworzenia jednokierunkowych sieci wielowarstwowych.

Sztuczne sieci neuronowe mają szereg zastosowań praktycznych, takich jak: zadania klasyfikacyjne, redukowanie zakłóceń[94], prognozowanie szeregów liczbowych oraz aproksymacja funkcji. Ostatnie z wymienionych zastosowań jest szczególnie istotne z punktu widzenia niniejszej pracy.

Masters [Masters1993] rozważa kilka szczególnych przypadków zadania przybliżania funkcji. Jednocześnie podając przykłady, konfrontuje wykorzystanie sztucznych sieci neuronowych z alternatywnymi lub tradycyjnymi metodami.

Dwa ze wskazanych we wspomnianej książce przypadków zadania przybliżania funkcji są bezpośrednio związane z poruszanym w niniejszej pracy zagadnieniem. Mianowicie są to modelowanie odwrotne oraz regresja wieloraka. W ramach modelowania odwrotnego rozważany jest przypadek, gdzie oryginalny model jest tradycyjny oraz stochastyczny. Sztuczne sieci neuronowe sprawują się dobrze w obu wariantach przy założeniu dostarczenia zbioru treningowego o odpowiedniej wielkości. Liczebność tego zbioru zdaniem autora „zależy od problemu i trudno odpowiedzieć z góry bez starannego przetestowania sieci", jaka powinna być. Z kolei w przypadku regresji wielorakiej Masters zauważa, że podejście statystyczne obarczone jest wadą związaną z liniowością relacji pomiędzy zmiennymi niezależnymi a zależnymi. W szczególnych przypadkach możliwa jest linearyzacja, niemniej w ogólności przy zależnościach nieliniowych metody statystyczne nie dają dobrych efektów. Stosowanie sieci połączeń

[94] Autoasocjacja lub wzmacnianie sygnałów.

funkcjonalnych [Pao1989] polecane jest dla modeli, w których znany jest przynajmniej przeważający charakter nieliniowości[95]. W przypadkach, gdy wiedza a priori o modelu nie istnieje, dobre rezultaty dają tradycyjne sieci z ukrytymi warstwami.

2.4.4 Drzewa decyzyjne

Pojęcie drzewa decyzyjnego odnosi się po pierwsze do stosowanego od lat 60. narzędzia wspierającego analizę decyzji, przede wszystkim w biznesie [Magee1964, Hespos1965]. Po drugie - do stworzonej na bazie tego narzędzia metody wykorzystującej drzewa decyzyjne jako modele predykcyjne.

Drzewa decyzyjne tworzą struktury zbliżone do diagramów przepływu. Każdy wierzchołek reprezentuje uszczegółowienie klasy wyniku, podczas gdy krawędzie obejmują wynik testu przeprowadzanego na elementach wierzchołka. Drzewa decyzyjne dobrze radzą sobie z zadaniami klasyfikacji[96], ale także można stosować je do zadań regresji[97].

Istnieje szereg metod uczenia drzew na podstawie danych złożonych z wektorów zmiennych niezależnych oraz wartości zmiennej zależnej. Do najpopularniejszych z tych metod zaliczyć można: algorytm C4.5 [Quinlan1996] i kontynuatorów, CART [Breiman1984], CHAID [Kass1980] czy CIT [Strobl2009].

[95] Pozwala to wyeliminować warstwy ukryte sieci i może spowodować istotny wzrost tempa uczenia. Z drugiej strony upodabnia metodę zastosowania sztucznej sieci neuronowej o takiej konstrukcji do specjalnego zastosowania regresji wielomianowej.
[96] Mowa wówczas o drzewach klasyfikujących. W takich przypadkach zakłada się, że zmienna wynikowa (zależna) ma charakter dyskretny.
[97] Wówczas mowa jest o drzewach regresji, a zmienna wynikowa (zależna) ma charakter ciągły.

3 Modele wyceny produktów ubezpieczeniowych

3.1 Produkt ubezpieczeniowy i jego charakterystyka

W literaturze marketingowej panuje pogląd, iż produkt stanowi w istocie złożenie trzech warstw. Warstwami tymi według Kotlera są: rdzeń produktu, produkt podstawowy oraz poszerzony [Kotler2012]. Dobrym odzwierciedleniem idei warstwowości w przypadku ubezpieczeń jest koncepcja prezentowana przez Hallera[98]. W koncepcji tej produkt ubezpieczeniowy składa się z produktu głównego, świadczeń podstawowych oraz świadczeń rozszerzonych. Warstwy te, co prawda w luźny sposób, ale odnieść można do wcześniej przedstawionej koncepcji.

Produktem głównym w przypadku ubezpieczenia jest ochrona ubezpieczeniowa, czyli „gotowość przejęcia przez zakład ubezpieczeń materialnych skutków realizacji ryzyk"[99]. Ochrona ta świadczona jest na zasadach oraz w zakresie ustalonym przez warunki umowne, zapisane w dokumentach towarzyszących zawarciu umowy ubezpieczenia[100].

Na świadczenia podstawowe, stanowiące kolejną warstwę w omawianym schemacie produktu, składa się szereg procesów i zasobów oddziałujących w sposób istotny na przepływ informacji oraz rozpowszechnienie wiedzy związanej z produktem głównym. Procesy i zasoby składające się na tę warstwę produktu mają charakter w części zadań koniecznych do zaistnienia produktu głównego, ale także wpływają na postrzeganie tego produktu w kategoriach zwiększonej wartości dodanej[101] przez nabywcę.

Warstwa trzecia ma charakter zbliżony do warstwy drugiej. Podstawowym wyróżnikiem zbioru świadczeń rozszerzonych, odróżniającym je od świadczeń zaliczanych do warstwy podstawowej, jest fakt braku bezpośredniego powiązania[102] tych pierw-

[98] [Haller1998], str. 561.
[99] [Handschke2000], str. 49.
[100] Ogólne warunki ubezpieczenia, tekst umowy, polisa.
[101] np. zwiększają wygodę, redukują czas etc.
[102] Przez brak bezpośredniego powiązania rozumieć należy tutaj także sytuację, w której cele ochronne są realizowane, ale za pomocą odrębnych narzędzi, np. zaliczających się do całościowego paradygmatu zarządzania ryzykiem.

szych z celami stawianym przed produktem podstawowym, czyli ochroną ubezpieczeniową.

3.1.1 Cechy produktu ubezpieczeniowego w procesie sprzedaży

Ubezpieczenia stanowią specyficzną grupę produktów. Specyfika ta wywodzi się z szeregu cech, które tradycyjnie przypisuje się usługom. Cechami tymi są[103]: niematerialność, różnorodność, nietrwałość.

Przez większą część cyklu życia produktu ubezpieczenie ma charakter niematerialny. Oznacza to, że istnienie ubezpieczenia, zawarcie i obowiązywanie umowy, na mocy której świadczona jest ochrona ubezpieczeniowa, sprowadza się w przeważającej mierze do procesów generowania oraz wymiany informacji pomiędzy kontrahentami.

Różnorodność jest cechą świadczącą o dużym zróżnicowaniu poszczególnych produktów w ramach całej oferty. Oznacza ona również, że ocena użyteczności produktu przez konsumenta zależy od szeregu czynników, w tym czynników obiektywnych oraz subiektywnych. Czynnikami obiektywnymi są przede wszystkim składowe temporalne, lokalizacyjne oraz metodyczne tworzące razem kontekst świadczenia usługi.

Przyjmuje się, że w przypadku usług, w odróżnieniu od namacalnych towarów, nie istnieje możliwość wytwarzania i przechowywania produktu w celu późniejszej odsprzedaży. Zjawisko takie określać można mianem nietrwałości lub niemożliwością magazynowania.

3.1.2 Marketing produktu ubezpieczeniowego

Marketing produktu ubezpieczeniowego to szereg działań i czynności zmierzających do aktywizacji popytu na oferowane produkty, sprzedaży tych produktów oraz zapewnieniu na odpowiednim poziomie obsługi posprzedażowej. Procesy marketingowe na etapie sprzedaży mają za zadanie zapewnienie odpowiedniego wsparcia dla przepływów pojawiających się w trakcie oferowania i realizacji ochrony ubezpiecze-

[103] W literaturze wymienia się także inne cechy, takie jak: „brak możliwości nabycia na własność” lub „substytucyjność”. Por. [Daszkowska1997], s. 17.

niowej. Dodatkowo mają one na celu także eliminację niezgodności, jakie pojawiają się przed lub po zawarciu umowy ubezpieczenia.

Istnieją trzy grupy strumieni zasobów związanych bezpośrednio z rozpoczynaniem i świadczeniem ochrony ubezpieczeniowej. W kolejności ich znaczenia są to: strumień informacyjny, strumień finansowy, strumień rzeczowy.

Dodatkowy podział, który wyłania się w związku z przepływami zasobów w ramach przytoczonych strumieni, dotyczy kwestii konstytutywności transferowanych zasobów względem wytworzenia ochrony ubezpieczeniowej. W tym przypadku wyróżnić można zasadniczo dwa typy przepływów: obligatoryjne oraz nieobligatoryjne. Różnica pomiędzy obydwoma polega na tym, że zasoby będące przedmiotem przepływów obligatoryjnych mają charakter konstytutywny dla ubezpieczenia, tj. ich zaistnienie jest niezbędne, aby ubezpieczenie mogło zacząć funkcjonować i pełnić swoją rolę.

Przepływy informacyjne mają zdecydowanie największe znaczenie spośród wszystkich strumieni zasobów towarzyszących dystrybucji ubezpieczenia. Mimo iż obydwie strony umowy nie posiadają początkowo dostatecznie precyzyjnej informacji o sytuacji i ofercie drugiej strony, przyjmuje się, że to zakład ubezpieczeń stoi na uprzywilejowanej pozycji, jeśli chodzi o przewagę negocjacyjną w zakresie omawianego zasobu. Zjawisko to nosi nazwę asymetrii informacji. Określone instytucje oraz wymogi prawne, a także względy praktyczne zmierzają do tego, aby w pewnym stopniu ową asymetrię minimalizować[104]. Jednak w zasadzie nigdy nie dochodzi do sytuacji, w której zjawisko to zostałoby zniwelowane całkowicie[105].

Strumień finansowy ma mniejszą wagę od wcześniej omówionego strumienia informacyjnego. Mimo to, przynajmniej w ograniczonym zakresie, wystąpienie tego strumienia także warunkuje funkcjonowanie ochrony ubezpieczeniowej. Strumień ten obejmuje obligatoryjne przepływy środków płaconych przez nabywcę ubezpieczenia w ramach ustalonej ceny, jaką jest składka. Warunki zapłaty składki mogą mieć bardziej skomplikowany przebieg w przypadku, gdy warunki umowy ubezpieczeniowej przewidują zapłatę w ratach lub odroczenie świadczenia w czasie.

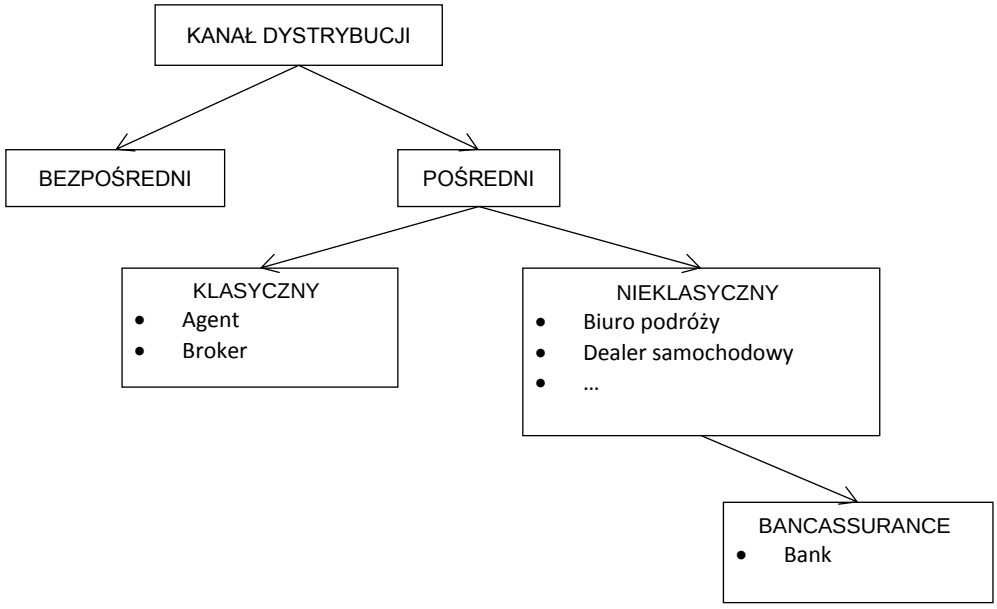
[104] Stanowi bowiem ona zjawisko niepożądane.
[105] Jest to m.in. jedna z przyczyn istnienia ról agenta czy brokera w procesach sprzedażowych.

Wystąpienie przepływów w drugą stronę, tj. od asekuratora do uposażonego, zależne jest od warunków świadczonej ochrony, w tym zajścia w określonym czasie zdarzenia powodującego straty na przedmiocie ubezpieczenia oraz metody restytucji. Jeżeli restytucja ma charakter pieniężny, w terminach przewidzianych przez regulacje umowne lub prawne nastąpić winien transfer odszkodowania.

Najmniejsze znaczenie w ramach omawianych procesów sprzedażowych mają transfery zachodzące w związku ze strumieniem rzeczowym.

3.1.3 Znaczenie kanałów marketingowych on-line

Zakres tradycyjnych definicji pojęcia kanału sprzedażowego nie obejmuje z dostateczną precyzją zmian, jakie wprowadził w technikach zbytu w ostatnich latach postęp technologiczny, a w szczególności intensywny rozwój technologii informatycznych i mobilnych. Ze względu na usługowy charakter ubezpieczeń, formuły definicyjne wymagają dodatkowej elastyczności, która wyklucza koncentrację na przejawach związanych wyłącznie z fizycznym przemieszczaniem dóbr w ramach kanału.



Rysunek 7. Podział kanałów dystrybucji ubezpieczeń
Źródło: [Kaczała2006]

W celu wyeliminowania nieporozumień co do faktycznych form kanałów dystrybucji w stosunku do ewentualnych ich odmian, wprowadza się dodatkowe pojęcie instrumentu sprzedażowego, mającego charakter uszczegóławiający sposób funkcjono-

wania danego kanału przez określenie szczególnych środków czy narzędzi wykorzystywanych do realizacji poszczególnych celów marketingowych[106].

Nie istnieje jednoznaczna konwencja co do klasyfikacji internetu (sieci WWW) jako kanału bądź instrumentu marketingowego. Podczas gdy niektórzy autorzy wskazują jednoznacznie na cechy upodabniające procesy sprzedażowe wykorzystujące internet do innych procesów związanych z poszczególnymi instrumentami, część organów nadzorczych rynku ubezpieczeniowego traktuje sprzedaż wykorzystującą nowoczesne technologie informatyczne jako pełnoprawny, odrębny kanał marketingowy i dystrybucyjny.

W dyskusji nad kwestią zaliczenia internetu do zakresu jednego z dwóch przytoczonych pojęć, poza szczegółowym określeniem do jakiej definicji będziemy się odnosić, należy także przeanalizować konkretny przypadek procesu sprzedażowego. Wykorzystanie internetu do dystrybucji dóbr i korzystanie z niego w przypadku dystrybucji ubezpieczeń może mieć bowiem mocno zróżnicowany zakres[107]. W najbardziej ograniczonym przypadku sieć WWW może być jedynie sposobem na jednostronną komunikację lub reklamę produktu. W najbardziej zaawansowanym wypadku – w zasadzie całość procesów marketingowych wraz z przekazem środków pieniężnych i przesłaniem podpisanych cyfrowo dokumentów może odbyć się wyłącznie w drodze elektronicznej wymiany danych. Przyjąć można punkt widzenia, zgodnie z którym w przypadkach zbliżonych do drugiego z przedstawionych skrajnych rozwiązań, do czynienia mamy raczej z kanałem dystrybucji, nie zaś czystym instrumentem komunikacji.

Zastosowanie zaawansowanych technologii informatycznych oraz historycznie nieco później mobilnych spowodowało szereg zmian, jakie zaszły w obszarze podmiotów uczestniczących w procesach dystrybucji. Po pierwsze, stworzyło okazję do pojawienia się na rynku ubezpieczeniowym nowych grup podmiotów[108]. Po drugie, w przypadku niektórych rodzajów ubezpieczeń zmniejszyło użycie tradycyjnych kanałów i instrumentów, a co za tym idzie, intensywność operacji wykonywanych przez

[106] Oczywiście dany instrument nie jest kanałem, mimo że czasem może być w ten sposób mylnie zakwalifikowany.
[107] Powiązaną tematycznie typologię prezentujemy w podrozdziale 4.1.
[108] Brokerzy informacji, portale porównujące oferty, wirtualni pośrednicy etc.

uczestników tych kanałów. Wreszcie aktywizowało podmioty do tej pory nieaktywne lub mało aktywne w powstających kanałach dystrybucji – mowa tutaj o ekspertach i specjalistach oraz zespołach zajmujących się wykorzystaniem zaawansowanych technologii do prowadzenia procesów sprzedażowych[109]. Dodatkowym elementem zmiany stała się możliwość większego wyodrębnienia na zewnątrz poza główną strukturę organizacyjną pewnych procesów[110]. W końcu nowym zjawiskiem stało się pojawienie nowych kategorii podmiotów, będących tworami wyłącznie wirtualnymi[111].

Decyzje w zakresie zarządzania kanałami sprzedaży obejmują takie działania jak: utrzymywanie, rozwój, tworzenie nowych oraz zamykanie nieperspektywicznych kanałów. Z każdym z utrzymywanych przez przedsiębiorstwo kanałów związany jest szereg parametrów, co stanowi o konieczności zarządzania kanałami sprzedaży. Wśród parametrów wyróżnić należy efektywność sprzedaży stanowiąca pochodną relacji kosztów utrzymania i prowadzenia operacji w ramach danego kanału i przychodów przez ten kanał przynoszonych. Z efektywnością kanału pośrednio powiązane są jego cechy charakterystyczne. Poszczególne typy kanałów w powiązaniu z wykorzystywanymi instrumentami dystrybucji są zróżnicowane pod względem swoich właściwości, a co za tym idzie, w danym otoczeniu rynkowym prezentują pewne przewagi lub słabości w stosunku do innych kanałów[112].

Dystrybucja produktów i usług z wykorzystaniem internetu wymaga uwzględnienia następujących elementów [Papazoglou2006]:

- potrzeby transportowe związane ze sprzedażą danego dobra,
- ograniczenia kulturowe, technologiczne i psychologiczne nabywców,
- bezpieczeństwo transakcji i kompletność oraz wiarygodność przekazywanych informacji,
- możliwe czynniki motywujące i zniechęcające,

[109] Infrastruktura teleinformatyczna, sprzęt, oprogramowanie, dedykowane systemy sprzedażowe, administracja systemów informatycznych.
[110] Outourcing usług IT, pójście w kierunku wirtualizacji organizacji.
[111] Mowa o agentach działających na rynkach elektronicznych. Agenty takie, co prawda do tej pory odgrywają marginalną rolę na rynkach ubezpieczeń, ale już niekoniecznie na rynkach związanych z zabezpieczeniem ryzyka – vide np. fundusze hedgingowe wykorzystujące mechanizmy algorytmicznego inwestowania na rynkach pochodnych papierów wartościowych i giełd towarowych.
[112] W gruncie rzeczy poszczególne kanały marketingowe z punktu widzenia organizacji mają charakter substytucyjny. Oznacza to, że przedsiębiorstwo powinno zatem optymalizować portfel kanałów uwzględniając ich efektywność i właściwości a także perspektywy sprzedaży w przyszłości.

- skłonność klientów do pełnej realizacji transakcji,
- zdolność do budowania bardziej długotrwałych relacji.

Lista wymienionych czynników nie ma charakteru kompletnego. Wszystkie argumenty, również tutaj nie uwzględnione, są rozpatrywane każdorazowo podczas formułowania i realizacji strategii marketingowej w różnych organizacja. Dzieje się tak w szczególności również w zakładach ubezpieczeń.

3.2 Metody wyceny ryzyka, obliczanie składki i konstrukcja systemów taryf

Metody tworzenia produktów ubezpieczeniowych w poszczególnych działach ubezpieczeń różnią istotnie. Wspólny jest niewątpliwie ogólny schemat przygotowania wyceny: zebranie wstępnych danych, kalkulacja składki netto, uwzględnienie doliczeń i odliczeń, kalkulacja składki brutto, obliczenie rezerw. Dobre omówienie metod stosowanych w przypadku ubezpieczeń na życie można znaleźć w szeregu pozycji, m.in. [Błaszczyszyn2004, Gerber1997]. Dalsze omówienie dotyczy przede wszystkim ubezpieczeń majątkowych.

3.2.1 Równowaga finansowa jako podstawowa przesłanka kształtowania cen ubezpieczenia

Podstawowym założeniem gospodarki finansowej zakładu ubezpieczeń, jak w każdym innym przedsiębiorstwie, jest założenie, że cena produktu powinna odzwierciedlać koszty poniesione w związku z jego wytworzeniem i sprzedażą [Jaworski2010]. Nadto winna także zawierać zakładaną marżę. Obrazuje to formuła:

$$P = C + Z, \tag{2}$$

gdzie:

- P – cena,
- C – koszt jednostkowy,
- Z – zysk jednostkowy.

Jedna z najbardziej fundamentalnych cech wyróżniających towarzystwa oferujące ubezpieczenia od innych przedsiębiorstw pojawia się już na tym poziomie kalkulacji. W przypadku przeważającej części dóbr i usług, koszt ich wytworzenia jest znany jeszcze przed momentem sprzedaży. Jak można się domyślić, taka zależność przyczy-

nowo-czasowa nie zachodzi dla produktów ubezpieczeniowych. Konsekwencją takiego stanu rzeczy jest to, że w przypadku produktów nie będących ubezpieczeniami względnie łatwo jest wyznaczyć poziom cen, dla których zrealizowany zostanie na określonym poziomie zysk jednostkowy. W przypadku ubezpieczeń natomiast proces ustalania cen, po których sprzedaż powinna się odbywać, jest zdecydowanie bardziej skomplikowany. Co więcej, w przypadku konkretnych, pojedynczych sztuk produktu, nie gwarantuje osiągnięcia zamierzonego celu ekonomicznego[113].

Ubezpieczenia są szczególnym rodzajem produktu finansowego ze względu na swoją konstrukcję w formie umowy cywilnoprawnej, w której jedna strona zobowiązuje się do zrealizowania określonego świadczenia w przyszłości w razie zajścia danych zdarzeń w umówionym czasokresie[114]. Przy czym zarówno sam fakt zajścia zdarzenia lub zdarzeń oraz związane z tym wielkości świadczeń nie są w momencie realizacji sprzedaży znane.

W rezultacie równaniu (2) można nadać nową formułę oraz interpretację. W takim ujęciu podstawowa zasada gospodarowania w ramach podmiotu gospodarczego w przypadku rynku ubezpieczeniowego przybiera następującą formę:

$$S_{brutto} = St + K_{lsz} + W_{UW} + P_{UW}, \tag{3}$$

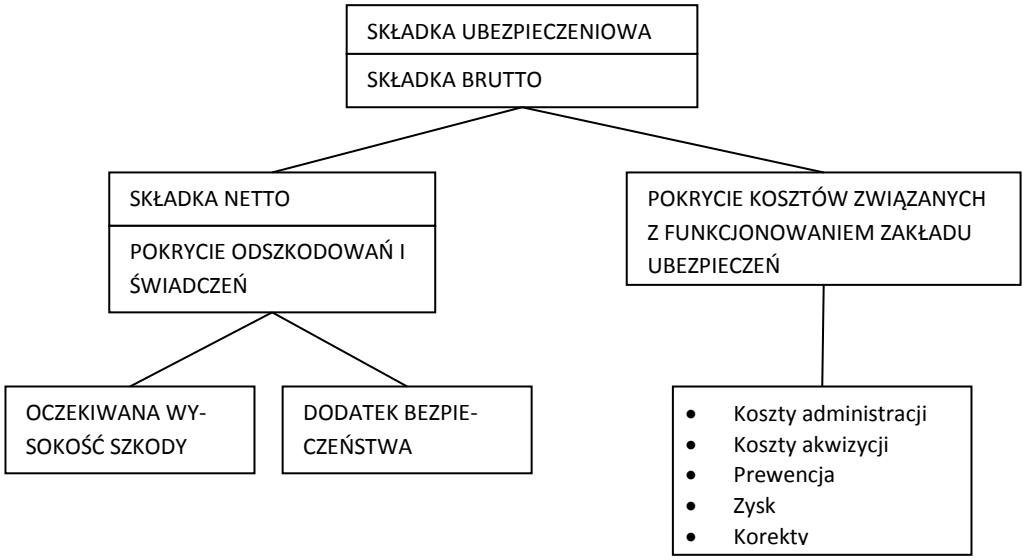
gdzie:

- S_{brutto} – składka brutto,
- St – bezpośrednie koszty likwidacji szkód (związane z ryzykiem),
- K_{lsz} – pośrednie koszty likwidacji szkód i koszty administracyjne,
- W_{UW} – koszty akwizycji (sprzedaży i pozyskania klientów),
- P_{UW} – przychody ze sprzedaży[115].

Szukając analogii pomiędzy równaniem (2) i (3) widać, że odpowiednikiem ceny w przypadku ubezpieczeń jest składka, natomiast przez koszt rozumieć należy sumy wydatków: z tytułu wypłat świadczeń i odszkodowań, a także na obsługę roszczeń, innych strat oraz pozostałych kosztów pozyskiwania polis[116]. Stwierdzić ponadto nale-

[113] Zakładana rentowność może pojawić się na poziomie agregacji nie zaś poszczególnych produktów.
[114] K.c. art. 805.
[115] Przychód ten jest różnicą pomiędzy wpływami i wydatkami z tytułu zawierania polis i jest analogiczny do zysku otrzymanego w innych sektorach.
[116] na które składają się koszty administracyjne, koszty akwizycji klientów, wydatki na działania prewencyjne, zyski oraz korekty – w tym m.in. inflacyjną, reasekuracyjną, z tytułu dochodów inwestycyjnych.

ży, że zysk ze sprzedaży polis jest tylko jedną z kategorii zysków istniejących w zakładach ubezpieczeń[117]. Ogólny schemat budowy składki prezentuje rysunek.



Rysunek 8. Tworzenie składki w oparciu o koszty zakładu ubezpieczeń
Źródło: **[Kowalczyk2006]**

Konieczność uwzględnienia fundamentalnych zasad gospodarowania przedsiębiorstwem w kalkulacji cen produktów ubezpieczeniowych znajduje swoje odzwierciedlenie w szeregu reguł, których zastosowanie wymusza zachowanie oraz przestrzeganie określonych relacji finansowych. Regułami tymi są:

1. reguła równowagi składki i świadczeń,
2. reguła proporcjonalności składek i świadczeń,
3. reguła równowartości składek i świadczeń[118].

Każda z powyższych reguł nakłada coraz to silniejsze ograniczenia na sposób kształtowania wielkości składek.

3.2.2 Miary ekspozycji na ryzyko

Do kryteriów właściwej miary ekspozycji na ryzyko zaliczyć należy: bezpośrednią proporcjonalność do oczekiwanej wielkości straty, praktyczność użycia oraz uwzględnienie wcześniejszych standardów miar.

Oczekiwana wielkość straty zależy od różnych czynników, które powinny zostać wykorzystane w postaci zmiennych ratingowych. Dzięki ich wprowadzeniu możliwe

[117] Patrz np. zysk z działalności finansowej.
[118] Zwana też „składką sprawiedliwą".

jest szczegółowe odzwierciedlenie różnic w możliwych poziomach ryzyka. Czynnik, który ujawnia najsilniejszy związek z wielkością strat, powinien być uznany za najwłaściwszy, aby stanowić podstawową miarę ekspozycji na ryzyko. Ponadto czynnik taki powinien mieć obiektywny charakter oraz powinien być względnie łatwy w pomiarze (ustaleniu rzeczywistej wartości) i weryfikacji.

Przykładowe miary ekspozycji na ryzyko

W tabeli 2 podajemy zestawienie przykładowych, często stosowanych miar ekspozycji dla różnych rodzajów ubezpieczenia. Wiedza o ich użyciu może stanowić istotną wskazówkę przy rekonstrukcji modeli wyceny składki.

Tabela 2. Przykładowe miary ekspozycji wg rodzajów ubezpieczenia
Źródło: [Werner2010]

Gałąź biznesu ubezpieczeniowego	Miary ekspozycji
Osobiste, samochodowe	Proporcja liczby ubezpieczonych samochodów w trakcie trwania polisy, wartość pojazdu
Dom, mieszkanie	Proporcja liczby ubezpieczonych nieruchomości w trakcie trwania polisy, wartość nieruchomości
Rekompensaty pracownicze	Wysokość wynagrodzenia, pozostałe informacje z listy płac
OC z tytułu prowadzenia działalności gosp. na zasadach ogólnych	Przychód, lista płac, powierzchnia, liczba jednostek produkcji
Majątek przedsiębiorstwa	Szacowana wartość majątku
OC z tytułu wykonywania zawodu lekarza	Liczba lat praktyki
OC z tytułu wykonywania zawodu	Liczba osób o danej specjalności (prawników, księgowych etc.)
Ruchomości osobiste	Wartość przedmiotu

3.2.3 Strona kosztowa

Równanie (3) przekształcić można do formy:

$$S_{brutto} = St + K_{lsz} + (K_{ST} + \%K_z \times S_{brutto}) + \%P_{UW} \times S_{brutto}, \quad (4)$$

gdzie:

- K_{ST} – koszty stałe,
- $\%P_{UW}$ - założona stopa przychodu ze sprzedaży,
- $\%K_z$ – stopa kosztów zmiennych,
- pozostałe oznaczenia jak w równaniu (3).

Na potrzeby obliczania składek zasadniczo wykorzystuje się trzy metody wyznaczania przypisu wydatków. Należą do nich:

1. metoda całkowitych wydatków zmiennych,
2. metoda projekcji opartej na składce [Schofield1998],

3. metoda projekcji opartej na ryzyku (polisie) [Childs1980].

Oczywiście przy stosowaniu powyższych metod generalnie oczekuje się, że kategoria wydatków ma charakter zmienny w czasie. Przede wszystkim wynika to z faktu istnienia ciągłej presji inflacyjnej, ale także innych czynników. Czasami analizy wysokości składek w portfelach ubezpieczycieli wykonuje się także w wartościach netto (koszty uwzględniają wówczas programy i obowiązujące umowy reasekuracji). Obecnie praktyka taka staje się coraz powszechniejsza ze względu na rosnące znaczenie reasekuracji oraz udział jej kosztów w strukturze wydatków ubezpieczycieli.

W szczególności w sytuacji podwyższonego kosztu kapitału wydatki na reasekurację uzasadniają włączenie ich udziału do obliczeń związanych z określaniem wysokości składek.

3.2.4 Strona przychodowa

W przypadku ubezpieczeń całkowity przychód stanowi suma przychodów ze składek oraz zysku inwestycyjnego wypracowanego z kapitałów ubezpieczyciela. Przedstawia to równość:

$$PC = P_{INW} + P_{UW}. \tag{5}$$

Istnieją dwa podstawowe źródła zysków inwestycyjnych: wolne środki zainwestowane oraz zysk inwestycyjny wypracowany przez fundusze pochodzące od ubezpieczonych. Z kolei przychód ze składek składa się z przepływów generowanych przez sprzedaż polis, co czyni go zbliżonym w swoim znaczeniu do przychodu ze sprzedaży w innych gałęziach gospodarki. Wielkość przychodów ze składek wyraża się wzorem:

$$P_{UW} = S_{brutto} - St - K_{lsz} - W_{UW}. \tag{6}$$

Ostatecznie udział w wydatkach oraz przychodach jest użyty do wyliczenia dwóch wskaźników, mówiących o relacjach kształtujących się w zakresie sprzedaży. Są to: dopuszczalny zmienny współczynnik strat (DZWS) oraz całkowity dopuszczalny współczynnik strat (CDWS). Oba parametry wyznaczone są na podstawie wzorów:

$$DZWS = 1.0 - \%K_z - \%P_{UW}, \tag{7}$$

$$CDWS = 1.0 - \%K_{ST} - \%K_z - \%P_{UW}, \tag{8}$$

gdzie:

- $\%K_{ST}$ – stopa kosztów stałych.

Pierwszy z wymienionych współczynników może być interpretowany jako procent każdej jednostki pieniężnej pochodzącej z opłaconej składki, który ma zostać przeznaczony na opłacenie oczekiwanych strat wraz z kosztami likwidacji szkód oraz przewidywanych kosztów stałych. Pozostała część natomiast stanowi procent każdej jednostki pieniężnej przeznaczonej na koszty zmienne oraz zysk.

Z kolei drugi współczynnik wyraża udział każdej jednostki pieniężnej pochodzącej z opłaconej składki, który ma zostać przeznaczony na opłacenie oczekiwanych strat wraz z kosztami likwidacji szkód. Pozostała część zostanie wydatkowana na koszty sprzedaży i pozyskania klientów oraz zysk.

3.2.5 Metody obliczania składki podstawowej

Jak zauważono w podrozdziale 3.2.1, składkę ubezpieczeniową zdekomponować można na składkę netto oraz jej zwiększenie składające się na składkę brutto. Klient płaci zawsze tę drugą składkę. Ze względów obliczeniowych oraz analitycznych po stronie zakładu ubezpieczeń często jednak operuje się na wielkości netto. Stąd wyróżnić można metody konstruujące samą składkę netto oraz takie, które w sposób kompleksowy dotyczą pełnej składki, tj. kwoty brutto. Metody związane z wyznaczaniem składek netto opierają się na różnorakich teoretycznych konstrukcjach, najczęściej wyrażanych w postaci zasad mówiących w jaki sposób powinna kształtować się obliczana wielkość. W szczególności są nimi[119]: zasada czystego zysku[120], zasada wartości oczekiwanej, zasada odchylenia standardowego oraz wariancji, zasada maksymalnej możliwej straty oraz metoda oparta na teorii użyteczności.

Z kolei jeśli chodzi o składkę brutto, to wyróżnić można dwa podejścia do wyznaczania całościowego poziomu składki: metoda czystej składki oraz metoda współczynnika szkodowości. Istotnym aspektem technik aktuarialnych jest także ciągły ich

[119] Sposoby wykorzystania podanych zasad przy kalkulacji poziomu składki zaprezentowane są dobrze w szeregu polskojęzycznych publikacji, np. w [Ronka-Chmielowiec2006].
[120] Inna nazwa to zasada równoważności składki.

rozwój m.in. przez włączanie nowych technik (symulacje [Salam2003], data mining[121]).

Metoda czystej składki

Pierwsza z metod uważana jest za bezpośrednie i proste ujęcie formuły określenia poziomu składki. Określa ona średni poziom wyznaczanej wartości, nie zaś zamianę w stosunku do aktualnego poziomu, jak to ma miejsce w przypadku drugiej metody. W ujęciu metody czystej składki, przeciętny poziom składki przypadającej na ryzyko wyznacza się za pomocą wzoru:

$$\overline{S_{\text{brutto}}} = \frac{[\overline{St} + \overline{K_{lsz}} + \overline{K_{ST}}]}{[1 - \%K_Z - \%P_{UW}]}. \tag{9}$$

Metoda współczynnika szkodowości

Metoda współczynnika szkodowości, w odróżnieniu od poprzednio przedstawionej, jest bardziej rozpowszechniona. Podejście to porównuje szacowany procent każdej jednostki pieniężnej ze składki potrzebnej na pokrycie przyszłych strat, wydatków związanych z regulacją strat i innych wydatków stałych z procentową wartością każdej jednostki pieniężnej, która jest dostępna dla opłacenia tych kosztów. Inaczej ujmując, metoda ta stanowi porównanie współczynnika sumy przewidywanych strat oraz wydatków związanych z regulacją strat i współczynnika stałych wydatków do współczynnika dopuszczalnej straty. Wyrażając powyższą relację wzorem, otrzymujemy następującą równość:

$$\Delta s_{brutto} = \frac{\left[\frac{\text{St} + K_{lsz}}{S'_{brutto}} + \%K_{ST}\right]}{[1 - \%\text{K}_\text{Z} - \%\text{P}_\text{UW}]} - 1. \tag{10}$$

Pomiędzy obydwoma przedstawionymi powyżej podejściami istnieją dwie istotne różnice. Pierwsza różnica polega na wykorzystanej w każdym z przypadków mierze strat. Wzór na wskaźnik strat bazuje na współczynniku strat (przewidywanej ostatecznej stracie i wydatkach związanych z regulacją strat przypadających na przewidywane składki przy obecnych cenach polis). Z kolei wskaźnik czystej składki bazuje na statystyce czystej składki.

[121] Metody data mining (drążenie danych) zostały już wspomniane wcześniej ze względu na zbieżności ich użycia w niniejszej pracy.

Kolejną istotną różnicą jest wynik obydwu podejść. W przypadku formuły wyliczającej współczynnik szkodowości, wynikiem jest stopa zmiany w stosunku do aktualnych stawek. Natomiast formuła wyliczająca czystą składkę podaje w rezultacie wartość składki rozumianą kwotowo. W rezultacie użycie pierwszej formuły możliwe jest tylko wówczas, gdy firma ubezpieczeniowa posiada już wyliczenie składek i podejmuje jedynie działania zmierzające do ich uaktualnienia. Nowo tworzony produkt ubezpieczeniowy wymaga opracowania z wykorzystaniem formuły na czystą składkę[122].

Podstawowy poziom składki jest zazwyczaj tak dopasowany, aby odpowiadał jednej z największych lub największej pod względem liczebności grup ryzyka. Dzięki temu najistotniejsze w budowie produktu ubezpieczeniowego analizy statystyczne przeprowadzane są na podstawie próby o dużej liczebności i przypuszczalnie stabilnych parametrach.

3.2.6 Taryfikacja jednowymiarowa

Ryzyko należy analizować przez pryzmat właściwych procedur taryfikacyjnych, których zadaniem jest grupowanie ryzyk o podobnym stopniu strat i określanie zróżnicowanych składek w celu uwzględnienia różnic w potencjale strat istniejących pomiędzy grupami.

Pierwszy etap klasyfikacji taryfikującej polega na określeniu, jakie kryteria umożliwiają skuteczną segmentację ryzyka na grupy[123]. Grupy takie powinny mieć wewnątrz podobną charakterystykę strat. Badane kryterium jest często określane jako zmienna taryfikacyjna. W niektórych przypadkach rozróżnia się zmienne: taryfikacyjne oraz polisowe[124]. Gdy populacja ubezpieczonych jest przydzielona do odpowiednich poziomów dla każdej zmiennej taryfikacyjnej, aktuariusz wyznacza relację danej taryfy w stosunku do poziomu podstawowego składki dla każdego taryfikowanego poziomu. W najprostszym wariancie relacja taka jest albo sumą (taryfikacja addytywna) albo iloczynem (taryfikacja multiplikatywna).

[122] Jest to zatem istotne ograniczenie dla ubezpieczycieli.
[123] Czasami używa się terminu klasa odnosząc się do grupy ubezpieczonych, którzy należą do tego samego poziomu dla każdej z kilku zmiennych taryfikacyjnych.
[124] Odpowiednio ang. rating variable oraz underwriting variable.

Dobór zmiennych taryfikacyjnych

Wskażemy obecnie na podstawowe kryteria określające dobór właściwych zmiennych taryfikacyjnych. Kryteria można podzielić na następujące kategorie [Finger2001]: statystyczne, operacyjne, społeczne, prawne.

Jeśli chodzi o kategorię kryterium prawnego, to konieczne jest, aby system klasyfikacji stawki był zgodny z obowiązującymi przepisami ustawowymi i wykonawczymi w każdym otoczeniu prawnym, w którym dana firma prowadzi biznes i oferuje ubezpieczenie.

Idealnym stanem rzeczy w przypadku opracowywania taryfikacji dla produktu ubezpieczeniowego jest sytuacja, w której ubezpieczyciel posiada lub może zebrać dane umożliwiające przetestowanie statystycznej istotności rozpatrywanego kandydata na zmienną taryfikacyjną. W takim przypadku wskazać można kryteria statystyczne, które powinny zostać wzięte pod uwagę w celu zapewnienia dokładności i wiarygodności potencjalnej zmiennej. Są nimi:

- istotność – zmienne taryfikacyjne powinny statystycznie istotnie różnicować ryzyko,
- jednorodność – poszczególne poziomy zmiennej taryfikacyjnej powinny odpowiadać oddzielnym grupom ryzyka, w których oczekiwane wysokości kosztów są bardzo podobne,
- wiarygodność – liczba jednostek ryzyka w każdej z wyszczególnionych grup musi być wystarczająco liczna lub charakterystyka tych ryzyk wystarczająco stabilna dla celów dokładnego oszacowania kosztów.

Osobnym zagadnieniem jest praktyczność zastosowania danej zmiennej w budowanym systemie taryfikacji. Zmiennej taryfikacyjnej można przyznać cechę praktyczności, jeżeli odpowiada ona trzem podstawowym aspektom:

- obiektywności – poziomy zmiennej taryfikacyjnej powinny być obiektywnie zdefiniowane i przejrzyście określone,
- ekonomiczności – koszty operacyjne związane z uzyskaniem informacji niezbędnych do prawidłowego klasyfikowania i oceniania danego ryzyka nie mogą być za wysokie,

- weryfikowalności – uzyskana wielkość zmiennej nie może podlegać wahaniom lub być narażona na manipulację w łatwy sposób przez uczestników kanału dystrybucji lub ubezpieczanego[125].

Oto kilka przykładów zmiennych według branży:

Tabela 3. Przykłady zmiennych taryfikacyjnych
Źródło: opracowanie własne na podstawie [Werner2010]

Gałąź biznesu ubezpieczeniowego	Zmienna taryfikacyjna
Osobiste ubezp. pojazdu	Wiek kierowcy, płeć[126], rocznik, historia wypadków
Ubezp. nieruchomości	Wartość, wiek nieruchomości, rodzaj budownictwa
Rekompensaty pracownicze	Kod w klasyfikacji zawodowej (np. wg KZiS)
OC z tytułu prowadzenia działalności gosp. na zasadach ogólnych	Klasyfikacja działalności, terytorium, limit odpowiedzialności
OC z tytułu wykonywania zawodu lekarza	Specjalność, terytorium, limit odpowiedzialności
Ubezp. zawodowego kierowcy	Klasyfikacja kierowcy, terytorium, limit odpowiedzialności

Jednowymiarowe metody taryfikacji

Metody jednowymiarowe są grupą metod upraszczających podejście do wyliczeń poziomów taryf. Obecnie mają one raczej znaczenie historyczne, ponieważ ustąpiły miejsca bardziej zaawansowanym metodom omówionym w kolejnych podrozdziałach. Mają one także znaczenie instruktażowe, szkoleniowe. Stosuje się je w szczególnych przypadkach. Następujące metody zostaną omówione: metoda czystej składki, metoda współczynnika strat oraz procedura minimalizacji nierównomierności rozkładu.

Metoda czystej składki

Metoda ta w podstawowej wersji polega na porównaniu wartości oczekiwanych czystych składek dla każdego z poziomów, które mogą być przypisane zmiennej taryfikacyjnej w celu ustalenia zachodzących relacji. Mając daną zmienną z_1 z określonym zróżnicowaniem taryfy dla każdego poziomu i oznaczonego jako z_{1i}, to stawka dla

[125] W przypadku kanału sprzedaży internetowej najczęściej sprowadza się to do akceptacji przez klienta przy zakupie regulaminu dającego prawo przedstawicielom ubezpieczyciela do inspekcji przedmiotu ubezpieczenia, która dokonywana jest w przypadku losowo wybranej grupy polis.
[126] Obecnie płeć nie może być zmienną taryfikacyjną na terenie UE.

każdego poziomu zmiennej z_1 (t_i) ustala się jako iloczyn stawki podstawowej (B)[127] i relacji wskaźnika (z_{1i}) :

$$t_i = z_{1i} \times B. \tag{11}$$

Czysta składka dla każdego poziomu jest oparta na statystycznym oszacowaniu kosztonośności każdego poziomu osobno i zakłada równomierny rozkład ekspozycji na ryzyko na wszystkich innych zmiennych taryfikacyjnych, co może skutkować tzw. efektem „podwójnego uwzględnienia".

Metoda współczynnika szkodowości

Główną różnicą między podejściem opisanym w ramach poprzedniej metody a metodą współczynnika szkodowości jest wykorzystanie aktualnych poziomów składek, a nie miary ekspozycji na ryzyko. Wzór wykorzystywany w metodzie współczynnika szkodowości wyraża się w sposób następujący:

$$\mathrm{WS} = \frac{\frac{(St + K_{lsz})_i}{S_{1,i}}}{\frac{(St + K_{lsz})_B}{S_{C,B}}}. \tag{12}$$

Ograniczenia w przedstawianej metodzie oraz założenia związane z danymi wejściowymi do przeprowadzenia procesu taryfikacji są podobne jak w przypadku poprzednio omówionego podejścia (np. niealokowalne koszty likwidacji szkód nie mogą być przydzielone do konkretnych klas). W omawianym ujęciu szczególnie istotne jest to, aby uzgodnić wartość zarobionej składki z aktualnym poziomem taryfikatora każdej klasy. Jest to wykonywane na dwa sposoby: albo za pomocą rozszerzenia ekspozycji[128], albo – jeżeli ograniczenia danych wykluczają użycie rozszerzenia ekspozycji – przez zastosowanie metody geometrycznej[129] na poziomie poszczególnych klas. Pierwszy sposób jest preferowany ze względu na większą dokładność.

Procedura minimalizacji nierównomierności rozkładu

Procedura minimalizacji nierównomierności rozkładu jest przykładem innego podejścia do zagadnienia klasyfikacji taryfikacyjnej. Wspólnym mianownikiem tej grupy

[127] Stawka ta wyznaczana jest za pomocą metody czystej składki omówionej w poprzednim podrozdziale.
[128] Uaktualnienia wartości z danych historycznych (np. składek) do obowiązujących wartości w celu zwiększenia liczby danych dla budowy statystyk.
[129] lub metody równoległoboku. Ma ona podobne zastosowanie co metoda rozszerzenia ekspozycji. Jest natomiast mniej precyzyjna, ponieważ zakłada grupowanie danych historycznych i przeliczanie tych wielkości za pomocą uśrednionych wskaźników.

algorytmów jest wykorzystanie standaryzowanych jednowymiarowych metod w ramach iteracyjnego procesu. W ramach procedury pierwszym krokiem jest zazwyczaj wybór odpowiedniej struktury taryfikacyjnej (przyjmowany jest model addytywny, multiplikatywny lub mieszany). Jednocześnie dobierana jest właściwa funkcja polaryzacji (np. zasada równowagi, najmniejszych kwadratów, χ^2, maksymalne nachylenie funkcji prawdopodobieństwa). Funkcja polaryzacji jest sposobem porównywania obserwowanych w ramach procedury statystyk strat (np. utraconych kosztów) do wskazanych statystyk strat i pomiaru ich wzajemnego niedopasowania. Obie strony tak skonstruowanej równości muszą być ważone ekspozycjami przynależnymi do poszczególnych grup taryfikacyjnych w celu dostosowania do nierównej mieszanki w portfelu danego ubezpieczenia. Termin „odchylenie minimalne" odnosi się do powszechnie stosowanej zasady równowagi[130]. Zasada ta nakłada wymóg, aby suma wskazanych ważonych czystych składek była równa sumie kosztów strat ważonych dla każdego poziomu dowolnej zmiennej taryfikacyjnej.

Istnieją artykuły, które opisują zastosowanie różnych procedur minimalizacji nierównomierności rozkładu. Szczegółowe, intuicyjne objaśnienia z przykładami zawarto w tekście [Feldblum2003].

Wady metod jednowymiarowych

Do wad podejść jednowymiarowych zaliczyć można to, że nie uwzględniają one dokładnie wpływu innych zmiennych taryfikacyjnych. Metoda czystej składki nie bierze pod uwagę korelacji ekspozycji z innymi zmiennymi taryfikacyjnymi. Jeśli algorytm taryfikacyjny operuje tylko na niewielkiej liczbie zmiennych, niedociągnięcie to może zostać załagodzone przez zastosowanie dwukierunkowej analizy lub dodatkowych ręcznie wprowadzonych zmian. W dzisiejszej praktyce aktuarialnej przy opracowywaniu taryfikatorów brane pod uwagę mogą być, w zależności od branży, dziesiątki lub setki kandydatów na zmienne taryfikacyjne. Powoduje to, że manualna regulacja jest co najmniej nieefektywna, a często wręcz niemożliwa.

[130] Jest to zasada analogiczna do reguły wspominanej w podrozdziale 3.2.1. Przy czym tam dotyczyła ona generalnie składki jako całości.

Wreszcie metody jednowymiarowe generalnie są czułe na składniki losowe w danych, co powoduje, że wyniki osiągane za ich pomocą obejmują zarówno rzeczywiste sygnały, jak i niepożądane szumy.

3.2.7 Metody wielowymiarowe kalkulowania taryf

W odróżnieniu od metod jednowymiarowych, podstawową korzyścią płynącą z zastosowania metod wielowymiarowych jest jednoczesne uwzględnienie wszystkich zmiennych taryfikacyjnych w ramach pojedynczej procedury obliczeniowej. Jak zauważono przy okazji krytyki metod opisanych w poprzednim podrozdziale, metody jednowymiarowe obciążone są dodatkowo nieuwzględnianiem korelacji miary ekspozycji pomiędzy zmiennymi taryfikacyjnymi. W przypadku obecnie prezentowanych metod obciążenie takie nie zachodzi.

Do innych istotnych przewag metod wielowymiarowych zaliczyć należy: filtrowanie danych z szumu, uwzględnienie interakcji pomiędzy poszczególnymi zmiennymi, tolerancję na specyficzne typy zmiennych[131], łatwe uzyskiwanie informacji diagnostycznej oraz przejrzystość[132].

Uogólniony model liniowy

Istotą techniki GLM[133] jest oszacowanie relacji zachodzącej między zmienną zależną (Y)[134] oraz wieloma zmiennymi objaśniającymi (predyktorami)[135]. W klasycznym modelu liniowym, z którego wywodzi się omawiana metoda, relacja taka przedstawiona może być jako suma wartości oczekiwanej zmiennej zależnej oraz czynnika losowego[136]. Przy czym wspomniana wartość oczekiwana traktowana jest jako kombinacja liniowa zmiennych objaśniających. Obrazuje to wzór:

$$Y = (\alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \cdots) + \varepsilon, \tag{13}$$

gdzie:

- $X_1, X_2, X_3, X_4, \cdots$ to kolejne zmienne objaśniające,
- $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \cdots$ to wartości oszacowanych parametrów modelu liniowego.

[131] Przez typ należy rozumieć definicję dziedziny wartości, jakie mogą być przypisane danej zmiennej. Przykładowo mogą to być zmienne enumeratywne, liczbowe całkowite, rzeczywiste etc.
[132] Z drugiej strony metody te mają bardziej charakter syntetyczny, podczas gdy metody jednowymiarowe są dalece bardziej analityczne, stąd lepiej nadają się dla celów demonstracyjnych czy edukacyjnych.
[133] ang. Generalized Linear Model.
[134] np. wskaźnik intensywności odszkodowawczej.
[135] np. wiek domu, kwota ubezpieczenia etc.
[136] Zwany również czynnikiem błędu (ϵ).

Zwykłe modele liniowe oparte są na szeregu założeń upraszczających – przede wszystkim: czynnik losowy ϵ ma rozkład normalny ze średnią równą 0 oraz stałą wariancję.

Rozwiązaniem zadania optymalizacji modelu liniowego jest znalezienie najlepszych estymatorów parametrów. Estymatory takie, zastosowane w modelu o określonej formie, generują dane przewidywane przez model z minimalnym błędem. Funkcją używaną do optymalizacji jest zazwyczaj funkcja sumy kwadratów błędów[137].

W praktyce w przypadku dużej liczby danych obserwacyjnych zawartych w zbiorach przeznaczonych do przeprowadzenia taryfikacji często wykorzystuje się różnorakie techniki numeryczne[138].

Uogólniona wersja modeli liniowych polega na usunięciu ograniczeń co do założeń normalności rozkładu oraz stałej wielkości wariancji. Ponadto wprowadza ona tzw. funkcję łącznikową, która pozwala określić związek pomiędzy zmienną zależną oraz kombinacją liniową zmiennych objaśniających. Wykorzystanie funkcji łącznikowych o różnorakiej postaci powoduje zasadniczą przewagę GLM w stosunku do podstawowej wersji modelu liniowego. W praktyce oznacza to, że zależność funkcyjna pomiędzy zmienną objaśnianą a zmiennymi objaśniającymi nie musi mieć charakteru liniowego[139]. Sama metoda uogólnionego modelu liniowego doczekała się licznych modyfikacji i ulepszeń, jak np. HGLM [Wolny-Dominiak2011]. Poszczególne warianty metody dodają lub zmieniają różne inne jej elementy lub założenia[140].

Bardziej szczegółowy opis metody GLM wraz z niezbędną teorią można znaleźć w części 1. monografii autorstwa Andersona [Anderson2007].

Techniki data mining

Do wielowymiarowych metod pomocniczych w zastosowaniach aktuarialnych zaliczyć można także m.in.: analizę czynnikową, analizę skupień, CART, MARS, sztuczne sieci neuronowe [Hastie2009].

[137] Inne funkcje są również możliwe do zastosowania. Np., [Lange1989], [Nievergelt1994].
[138] np. wielowymiarowa wersja algorytmu Newton-Raphsona.
[139] np. model GLM można wykorzystać do modelowania kształtowania się roszczeń ubezpieczeniowych dla celów taryfikacji; stosuje się wówczas logarytmiczną funkcję łączącą, która zakłada, że zmienne taryfikacyjne połączone są w sposób multiplikatywny.
[140] np. warunki przesunięcia, wagi priorytetów.

Analiza czynnikowa to narzędzie służące do eliminacji nadmiarowej liczby kandydatów na zmienne taryfikacyjne. Rezultatem zastosowania jest zredukowanie liczby zmiennych lub dziedziny przyjmowanych przez nie wartości. Dzięki temu rozrośnięta lista zmiennych skorelowanych, mających nikły wpływ na rezultaty modelu, jest eliminowana lub scalana do jednej lub kilku nieskorelowanych zmiennych zastępczych[141].

Analiza skupień jest narzędziem służącym do eksplorowania przestrzeni obiektów[142] w celu wykrycia podobieństw i różnic, a w konsekwencji – optymalizacji podziału tych obiektów na grupy. Podział odbywa się w taki sposób, że grupowane są obiekty maksymalnie podobne, tworząc większe, jednorodne skupiska, jednocześnie minimalizując podobieństwa pomiędzy całymi grupami. Grupowanie ryzyk jest istotnym elementem poprzedzającym proces taryfikacji, a także umożliwiającym dywersyfikację jego rezultatów. W praktyce analiza skupień jest najczęściej stosowana w taryfikacji na podstawie kryteriów geograficznych, gdzie analiza zaczyna się na niskim poziomie granulacji jednostek geograficznych i dąży się do jego zwiększenia.

CART[143] jest metodą optymalizacji klasyfikacji przez uczenie algorytmów funkcjonujących na podstawie drzew decyzyjnych.

MARS[144], Wielowymiarowa Regresja Adaptacyjna Funkcji Sklejanych, to nieparametryczna metoda regresji stanowiąca rozszerzenie modeli liniowych w taki sposób, że pozwala na automatyczne uwzględnienie nieliniowości w modelowanym systemie. Rezultat działania stanowi funkcja będąca złożeniem równań regresji liniowej dopasowanych do odcinków oddzielonych odpowiednio dopasowanymi punktami granicznymi. Technika ta jest powszechnie stosowana w celu wyboru granic przedziałów przy kategoryzacji zmiennych ciągłych.

3.2.8 Rozszerzanie analizy wieloczynnikowej o dane zewnętrzne

Rozwój wielowymiarowych metod taryfikacji spowodował wzrost zainteresowania firm ubezpieczeniowych w wykorzystaniu zewnętrznych źródeł danych, zarówno do rozszerzenia posiadanych systemów taryfikacyjnych, jak i przy tworzeniu takich

[141] Stanowią one kombinację liniową zmiennych zastępowanych w modelu.
[142] np., ryzyk.
[143] ang. Classification and Regression Trees.
[144] ang. Multivariate Adaptive Regression Spline.

systemów na potrzeby nowo wprowadzanych produktów. Szczególnymi obszarami zainteresowania w takich przypadkach są dane:

- geo-demograficzne (np. gęstość zaludnienia na danym obszarze, średnia długość posiadania domu na danym obszarze)[145],
- meteorologiczne (np. liczba dni z temperaturą poniżej 0 stopni Celsjusza, średnie ilości opadów na danym obszarze),
- dane dot. nieruchomości (średnia wartość, powierzchnia, wyposażenie jednostek straży pożarnej w sąsiedztwie),
- inne dane statystyczne (osobowe lub dot. podmiotów gospodarczych).

Analiza terytorialna

W przypadku wielu rodzajów ryzyk, różne miary je charakteryzujące, takie jak szkodowość czy prawdopodobieństwo wypadku ubezpieczeniowego, są silnie zależne od otoczenia przedmiotu ubezpieczenia. Charakteryzowanie tego otoczenia za pomocą szeregu zmiennych bywa trudne lub nieefektywne. W takich wypadkach lepiej jest zbiorczo traktować te charakterystyki jako funkcję lokalizacji przedmiotu ubezpieczenia.

Lokalizacja jest czynnikiem o szczególnej wadze w przypadku charakterystyki wielu rodzajów ryzyk. Zatem położenie przedmiotu ubezpieczenia jest powszechnie wykorzystywaną zmienną taryfikacyjną, której wykorzystanie niesie jednak za sobą określone praktyczne wyzwania.

Po pierwsze, wyzwaniem jest właściwe zdefiniowanie zmiennej taryfikacyjnej[146], obejmujące typ, zakresy, dziedzinę wartości, ograniczenia etc. W praktyce aktuarialnej, ze względu na postulat praktyczności użycia zmiennych, terytoria traktowane

[145] Zwrócić uwagę warto tutaj na szereg prac, przykładowo Gobble i Windeler [Gobble2003] wskazują na znaczący postęp w systemach oceny ryzyka katastrof ze względu na napływ danych z nowej generacji systemów informacji geograficznej, cechujących się dużo wyższą jakością, w tym ich dokładnością. Spowodowało to, zdaniem autorów, znaczący wzrost możliwości zróżnicowania ocen ryzyka dla celów ubezpieczeniowych. Lane i Dennis [Lane1985] badają wpływ użycia różnych danych socjoekonomicznych oraz ich estymacji dla celów tworzenia profili klientów oraz ich oceny ze względu na możliwości prognozowania samych wypadków ubezpieczeniowych (bez względu na szkodowość). Dyskutowany jest także wpływ współzależności pomiędzy poszczególnymi elementami profilu..

[146] Lokalizacja może być opisana jako punkt na mapie lub regularny obszar albo administracyjnie wytyczone terytorium etc.

są jako zbiór małych jednostek geograficznych[147]. Poszczególnym jednostkom przypisane są modyfikatory taryfikacyjne.

Po drugie, istotnym problemem w analizie terytorialnej jest ograniczenie ilościowe danych przypadających na poszczególne jednostki terytorialne. Występuje tu prosta zależność, iż im drobniejsza jest granulacja jednostek, tym podzbiory danych są mniej liczebne[148]. W celu ominięcia tego niekorzystnego zjawiska stosowane są specjalne techniki wieloczynnikowe.

Po trzecie, wyzwaniem może być sugerowane już wcześniej silne skorelowanie lokalizacji z innymi potencjalnymi zmiennymi taryfikacyjnymi[149]. Silna korelacja sprawia, że tradycyjna analiza jednoczynnikowa staje się często bezużyteczna.

3.3 Źródła wiedzy dla ubezpieczeń

Przyjąć można, że dla większości typów biznesów ubezpieczeniowych, minimalny zestaw wiedzy niezbędny do obsługi procesów związanych z produktem ubezpieczeniowym stanowią [Werner2010]: opis reguł, opracowanie taryfikacyjno-składkowe, algorytm taryfikacyjny oraz wytyczne dotyczące polisowania.

Najczęstszym rozwiązaniem jest podział tak skategoryzowanej wiedzy na dwa typy dokumentów: podręcznik taryfikacji oraz, zazwyczaj odrębny, podręcznik sprzedaży polis. W takim przypadku podręcznik taryfikacji zawiera trzy pierwsze elementy wymienione na powyższej liście, natomiast podręcznik sprzedaży polis zawiera treści składające się na ostatnią pozycję.

Pierwszy rodzaj dokumentów zawiera przede wszystkim treść dotyczącą informacji jakościowej potrzebnej do zrozumienia i zastosowania ilościowych algorytmów taryfikacyjnych zawartych w dalszej części dokumentacji. Opis reguł zwyczajowo poprzedzony jest zbiorem definicji szczegółowo określających ubezpieczane ryzyka oraz istotne kwestie z nimi powiązane. Istotną częścią, obszerną objętościowo, są szczegółowe wytyczne określające zasady klasyfikacji ryzyk, poprzedzające zastosowanie algorytmu taryfikacyjnego.

[147] np. kody pocztowe, powiaty, bloki spisowe.
[148] Zależność ta nosi nazwę „wysokiej wymiarowości" (ang. high-dimensionality).
[149] np. wysokiej wartości domy często są zlokalizowane w tej samej dzielnicy lub wręcz budynku.

Numeryczne zestawienia niezbędne do wyznaczenia ostatecznego poziomu składki są zawarte w opracowaniu taryfikacyjno-składkowym. Opracowanie takie zestawia m.in. stawki bazowe, tabele modyfikatorów oraz wykazy opłat i prowizji. Stawki bazowe odzwierciedlają poziom kosztów ubezpieczenia dla odgórnie założonego przy tworzeniu modelu składkowo-taryfikacyjnego bazowego profilu ryzyka[150]. Taki profil ryzyka w większości przypadków odpowiada charakterystykom ryzyka, które są najczęściej obserwowalne. Ewentualnie może być także wynikiem działań marketingowych ubezpieczyciela.

W przypadku złożonych produktów ubezpieczeniowych, w których polisy chronią ubezpieczonego od zróżnicowanych ryzyk[151], składka obliczana jest oddzielnie.

Zasadniczo dla profili ryzyka różniących się od profilu bazowego także poziom składki jest inny od założonej stawki bazowej. Algorytm taryfikacyjny odpowiedzialny jest za właściwe zróżnicowanie stawek bazowych dla różnych charakterystyk ryzyka. Osiągnięte jest to przez zastosowanie specyficznych formuł matematycznych, a także wykorzystanie modyfikatorów o charakterze multiplikatywnym lub addytywnym. Ponadto ostateczna składka musi uwzględniać jeszcze doliczone koszty z tytułu pozyskania klienta oraz zawarcia i obsługi umowy ubezpieczenia. Koszty takie mogą być stałe lub zmienne – w zależności od sposobu ich naliczania i stopnia powiązania z poszczególnymi procesami biznesowymi w przedsiębiorstwie[152]. Zmienne koszty naliczane są jako odsetek od składki bazowej lub zindywidualizowanej[153].

Algorytm taryfikacyjny powinien mieć dostatecznie wysoki poziom szczegółowości, aby nie pominąć żadnych – nawet z pozoru nieistotnych – aspektów procedury. Wiedza zawarta w algorytmie obejmuje:

- kolejność uwzględniania zmiennych profilujących,
- sposób wpływu zmiennych taryfikacyjnych na poszczególne kroki wykonywania obliczeń,

[150] Stawka bazowa jest taryfą, która ma zastosowanie w przypadku bazowego wariantu ryzyka. Rzadko jest to stawka uśredniona.
[151] Dobrym przykładem takiego rodzaju produktu są osobiste ubezpieczenia samochodowe, gdzie standardowo oddzielnie ocenia się podstawowe ryzyko oraz odpowiadającą mu podstawową taryfę wraz z tabelami taryfikacyjnymi dla każdego wariantu ubezpieczenia.
[152] Kosztem stałym jest np. koszt wystawienia polisy.
[153] Przykładem może być prowizja agenta.

- ograniczenia i limity oraz sposoby postępowania w przypadku ich przekroczenia, nakładane na wartości zmiennych taryfikacyjnych lub wyników poszczególnych kroków działania algorytmu,
- szczegółowe aspekty związane z przeprowadzaniem operacji matematycznych na danych[154].

Wytyczne dotyczące polisowania

Wspomniane wytyczne dotyczące polisowania mają charakter decyzyjny i poprzedzają proces kalkulacji składki. Są one silnie uzależnione od kondycji i specyfiki firmy ubezpieczeniowej, ponieważ definiują zbiór, zakres oraz szczegółowy profil ryzyk, od jakich ubezpieczyciel zgadza się świadczyć ochronę. Jeżeli ryzyko według wytycznych jest ubezpieczalne, to dodatkowo mogą one kształtować (modyfikować) poszczególne kwestie związane z procedurą wyliczania składki.

Bazy danych i dane ubezpieczeniowe

Opracowywanie nowych produktów ubezpieczeniowych stanowi mniejszą część pracy aktuariuszy. Częstszym obowiązkiem jest analiza i uaktualnianie taryfikatorów dla istniejących produktów. W związku z potrzebą dość regularnego przeprowadzania takich analiz, częściej używane są wewnętrzne dane gromadzone przez firmę ubezpieczeniową lub też dane branżowe[155].

Inaczej dzieje się w przypadku opracowywania nowego produktu ubezpieczeniowego. Przy realizacji takiego zadania najczęściej wymagane jest wykorzystanie danych zewnętrznych związanych z tworzonym produktem lub też nabycie danych niepublicznych od innej instytucji lub firmy.

Analiza taryfikacyjna operuje zazwyczaj na dwóch rodzajach danych wewnętrznych (rzadziej zewnętrznych):

- dane o zagrożeniach – polisy, składki, liczby roszczeń,
- dane księgowe – zagregowane raporty o kosztach.

Kwestie jakości danych dla potrzeb aktuarialnych poruszane są w różnych opracowaniach, stanowią również przedmiot standardów [ASOP2004].

[154] np. sposoby przeprowadzania zaokrągleń wartości.

[155] Na marginesie rozważań nad modelowaniem danych dla ubezpieczeń, warto zwrócić uwagę na dwa dodatkowe konteksty: możliwości i zastosowanie komercyjnych systemów informatycznych [DFA2007], a także próby budowy pierwszych systemów semantycznych dla firm ubezpieczeniowych [Tomaszewski2009].

Analiza taryfikacyjna jest w gruncie rzeczy procesem kojarzącym ze sobą dwa zbiory informacji: informacja o zawieranych polisach, ocenie ryzyka oraz płaconej składce konfrontowana jest z informacją o roszczeniach i wypłatach odszkodowań wraz z innymi kosztami działalności. W rzeczywistości firm ubezpieczeniowych informacja taka często gromadzona jest w dwóch lub nawet większej liczbie oddzielnych baz danych.

Wykorzystanie danych zewnętrznych

Wykorzystanie danych zewnętrznych jest częstą koniecznością dla zakładów ubezpieczeniowych w sytuacji opracowywania nowego produktu lub wkraczania na nowy obszar rynku ubezpieczeniowego. Z drugiej strony uzasadniona możliwość użycia danych w postaci materiału uzupełniającego, pochodzących z obcych źródeł, pojawia się także w przypadku zmian lub uaktualnień produktów już znajdujących się w ofercie.

Najczęściej wykorzystywanymi źródłami danych zewnętrznych są dane statystyczne, zbierane zarówno przez instytucje publiczne, jak i prywatne, zagregowane dane ubezpieczeniowe, informacje o taryfikacji produktów konkurencyjnych, a także dane nie związane z działalnością ubezpieczeniową[156].

W powszechnym użyciu są także dane zbierane lub agregowane przez specjalnie w tym celu powołane organy, specyficzne dla rynku ubezpieczeń w zależności od branżowych regulacji w danym kraju lub regionie[157].

[156] np. dane branżowe w przypadku specyficznych produktów, przykładowo dot. rolnictwa, energetyki etc.
[157] W Polsce może to być np. CEPiK [CEPiK] oraz bazy danych OI UFG [OIUFG].

4 Portale oferujące produkty ubezpieczeniowe

4.1 Klasyfikacja portali oferujących ubezpieczenia

Na potrzeby pracy wyróżniamy trzy podstawowe klasyfikacje portali ubezpieczeniowych:

- wg mieszanego kryterium modelu biznesowego oraz podmiotowego,
- wg kryterium pełności oferty – stopnia automatyzacji procesu sprzedaży,
- wg kryterium technologii wykonania witryny - pokrywający się z przedstawioną w podrozdziale 2.1 klasyfikacją modelowych źródeł internetowych.

Podział ze względu na model biznesowy

Pierwsze z prezentowanych kryteriów dotyczy klasycznie wyróżnianych modeli biznesowych, typowych dla obrotu elektronicznego, lecz zastosowanych do rynku ubezpieczeń. Zróżnicowanie modeli biznesowych odbywać się może w trzech obszarach, tj. B2B, B2E oraz B2C[158]. W przypadku dystrybucji produktów ubezpieczeniowych, pierwsze dwa obszary uznać można za mało interesujące z powodu względnie zamkniętego schematu przepływów informacji.

W ramach obszaru B2C wskazać można w handlu elektronicznym produktami ubezpieczeniowymi następujące modele biznesowe oraz powiązane z nimi typy witryn:

- dystrybucja bezpośrednia przez witryny zakładów ubezpieczeń,
- dystrybucja pośrednia za pomocą witryn pośredników wyłącznych,
- dystrybucja pośrednia za pomocą witryn pośredników niezależnych,
- marketing produktów za pomocą punktów sprzedaży[159],
- marketing produktów za pomocą wortali tematycznych.

Model dystrybucji bezpośredniej przez witryny zakładów ubezpieczeń jest przypadkiem najbardziej oczywistym i zarazem szeroko rozpowszechnionym. Koszty utrzymania witryny są zdecydowanie niższe w porównaniu z wydatkami na niezbędną

[158] Powszechnie występujące akronimy oznaczające relacje biznesowe pomiędzy podmiotami gospodarczymi (B2B), podmiotem gospodarczym a podmiotem zatrudniającym (B2E) oraz konsumentem (B2C).
[159] PoS – ang. Point of Sale.

infrastrukturę w przypadku innych kanałów sprzedaży[160]. Mimo to liczba witryn, które przyporządkować można do tego modelu, jest ograniczona. Jest to wynikiem określonej liczby ubezpieczycieli funkcjonujących na rynku, która z kolei jest konsekwencją regulacji prowadzenia działalności[161]. W rozdziale 4.3 przytaczamy wyniki badań, z których wynika, że obecnie każdy zakład ubezpieczeń w Polsce posiada swoją witrynę. W praktyce jednak witryn takich jest więcej, ponieważ częstym przypadkiem jest rzeczywisty podział treści prezentowanych przez zakłady ubezpieczeń pomiędzy kilka wyspecjalizowanych portali. Wydzielenie to oznacza przede wszystkim funkcjonowanie witryn pod różnymi domenami, ale także często zróżnicowanie na poziomie szaty graficznej, metod prezentowania treści oraz technologii ich generowania. Abstrahując od podziału witryn odzwierciedlających nakaz separacji działalności w ramach działów[162], częstym zabiegiem jest tworzenie oddzielnie witryn przeznaczonych dla prezentowania informacji o podmiocie[163], natomiast prezentacja oferty w zakresie sprzedaży on-line wraz z jej realizacją prowadzona jest przez odrębny portal. Jeżeli istnieje także portal przeznaczony dla agentów, to również częstą praktyką jest wydzielenie go w formie niezależnej platformy webowej[164].

Z punktu widzenia ekstrakcji wiedzy, portale korporacyjne stanowią najbardziej wiarygodne źródło internetowe ze względu na zminimalizowany czas i zasięg[165] niezbędnego obiegu wiedzy w przypadku jakichkolwiek zmian w ofercie produktowej.

Drugi model, polegający na dystrybucji pośredniej przez witryny pośredników wyłącznych, obejmuje przede wszystkim proste strony internetowe tworzone przez agentów świadczących usługi w sferze niewirtualnej. Strony takie stanowią najczęściej formę wizytówek reklamowych, nie zawierają natomiast cenników, a tym bardziej interaktywnych kalkulatorów, albowiem mechanizmy takie są zbyt skomplikowane w utrzymaniu, a w szczególności kłóciłoby się to z naturą i interesem podmiotu

[160] Pomijamy tutaj istotny składnik kosztów jakim jest reklama witryny oraz domeny internetowej.
[161] Ustawa z dnia 22 maja 2003 r. o działalności ubezpieczeniowej (Dz.U. 2003 nr 124 poz. 1151).
[162] Co w praktyce oznacza dwa różne portale dwóch oddzielnych choć faktycznie powiązanych spółek. W zasadzie nie zdarza się, żeby spółka z jednego działu posiadała na swojej witrynie ofertę podmiotu operującego w dziale drugim. Często natomiast istnieją dowiązania pomiędzy portalami.
[163] ew. także opisu oferty i rozwiązań produktowych.
[164] Mowa jest tu o portalach wewnętrznych z dostępnym publicznie interfejsem użytkownika.
[165] Mierzony jako liczba instancji oraz agentów pomiędzy którymi dochodzi do wymiany informacji lub wiedzy. Każdy proces wymiany może prowadzić do zniekształceń lub uproszczeń.

utrzymującego stronę. W konsekwencji z punktu widzenia niniejszej analizy są to źródła o małej użyteczności.

W kolejnym modelu, jakim jest dystrybucja pośrednia za pomocą witryn pośredników niezależnych, mamy do czynienia z bardzo szeroką gamą przypadków. Jest to także znaczna grupa pod względem ilościowym[166]. Zasadniczo wyróżnić można w tym modelu dwa typy stron internetowych: strony multiagentów oraz witryny brokerów. Przez szeroką gamę przypadków rozumiemy tutaj witryny, które tematyką oraz wykonaniem przypominać mogą zarówno portale ubezpieczycieli (zarówno te z treściami prezentacyjnymi, jak też sprzedażowe) oraz takie, które będąc bardziej ubogie, zbliżają się do prostych stron agentów. Portale liczących się na rynku brokerów mają czasami własne kalkulatory ubezpieczeniowe[167].

Podobny stopień niejednorodności prezentują dwa ostatnie wyróżnione modele biznesowe. Zakwalifikować można tutaj heterogeniczny zbiór stron, poczynając od wortali tematycznych związanych z ubezpieczeniami przez portale z innych branż (turystyka, samochody), aż do witryn bankowych działających w ramach koncepcji bancassurance. Można także zaklasyfikować do tej grupy niektóre systemy porównujące oferty ubezpieczeniowe[168]. Ze względu na wewnętrzne zróżnicowanie, możliwość wyekstrahowania wiedzy z tych portali prezentuje także zdywersyfikowany poziom, który należy oceniać na zasadzie poszczególnych przypadków.

Podział wg kryterium pełności oferty

Kryterium to nawiązuje do stopnia zautomatyzowania procesu sprzedażowego. Zasadniczo wyróżnić można takie przypadki, jak:

1. wyłącznie prezentacja oferty, kontakt telefoniczny[169],
2. możliwość umówienia spotkania z agentem,
3. formularz ze zmiennymi taryfikacyjnymi wysyłany i przetwarzany manualnie,
4. kalkulator ubezpieczeniowy bez możliwości zakupu on-line[170],

[166] Interesujące statystyki na ten temat zamieszczono tutaj: http://www.abport.com.pl/multiagent-agencja-ubezpieczeniowa-knf-polska, odczytano 29-05-2015 r.
[167] np. https://cuk.pl/o_nas/o_cuk, odczytano 31-05-2015 r.
[168] Które technicznie nie pośredniczą w samym zawarciu umowy – czyli nie są przedstawicielami ani brokerami.
[169] Czasami w przypadku prostych systemów taryfikacji lub wręcz ich braku podawane są tabele z cennikiem.

5. porównanie różnych ofert[171],
6. pełen proces zakupu poprzez witrynę internetową[172],
7. pełen proces zakupu oraz obsługa posprzedażowa.

Cząstkową wartość, jeśli chodzi o zadanie ekstrakcji wiedzy, posiadają reprezentanci należący do trzeciego z wymienionych elementów, jednakże nasze zainteresowanie skupia się na stronach internetowych, jakie zaliczyć można do ostatnich czterech grup.

4.2 Charakterystyka sprzedaży ubezpieczeń przez internet

4.2.1 Portale produktowe zakładów ubezpieczeń

Portale produktowe należące do towarzystw ubezpieczeniowych są portalami oferującymi produkty zasadniczo tylko jednego ubezpieczyciela. Liczba produktów bywa zróżnicowana i zależy od zakresu oferty towarzystwa w ogóle, wahając się od jednego do kilku produktów.

Strona startowa portali ma za zadanie przedstawienie oferty produktowej w podziale na rodzaje ubezpieczeń, które odwiedzający może nabyć. Ekrany z formularzami wykonane są w różnorodnej technologii, ale z dominacją zaawansowanych technik omówionych w podrozdziałach 2.1.2-2.1.6.

Portale internetowe oferujące usługę wyceny ubezpieczenia mają zazwyczaj skomplikowaną, niehomogeniczną budowę. Cechuje je także wysoki poziom zaawansowania, jeśli chodzi o wykorzystywane technologie budowy oraz prezentacji treści, choć spotkać można również proste aplikacje internetowe. Z drugiej strony część tego typu witryn ma specyficzne zabezpieczenia przed niepowołanym dostępem lub próbami zbyt intensywnego dostępu do prezentowanych treści. Zabezpieczenia te podzielić można na dwa rodzaje: filtrowanie komunikacji oraz filtrowanie użytkownika. Pierwsza grupa to przede wszystkim zabezpieczenia serwera przed obsługą nadmiernej

[170] W niektórych przypadkach oferta nie jest podawana bezpośrednio jako rezultat działania zainteresowanego klienta lecz rezultat ma charakter odroczony – informacja zwrotna jest wówczas zazwyczaj wysyłana na wskazany email.
[171] Porównanie ofert ubezpieczycieli odbywać się może w dwóch wymiarach: cenowym oraz jakościowym (różnice w regulacjach OWU) oferty. W przypadku nielicznych portali porównujących produkty ubezpieczeniowe na polskim rynku zdecydowanie dominuje podejście pierwsze. Piszemy o tym szczegółowo w kolejnych podrozdziałach.
[172] W takim przypadku prawie zawsze oferowana jest pomoc asystenta lub możliwość kontaktu z agentem.

liczby połączeń z określonych węzłów sieci. Do grupy drugiej zaliczyć można m.in. mechanizmy typu CAPTCHA[173], konieczność rejestracji lub identyfikację za pomocą danych personalnych.

4.2.2 Portale porównujące ofertę

Idea portali porównujących oferty polega na agregacji informacji o produktach lub usługach dostarczanych przez różnych sprzedawców. Zestawienia tego typu mogą być szczególnie użyteczne dla potencjalnego klienta w przypadku produktów homogenicznych, tj. takich towarów, które ew. różnią się jedynie podstawowymi cechami, a różnice te zazwyczaj nie mają wpływu na cenę. W przypadku przenoszenia tej koncepcji na rynek ubezpieczeniowy, implementacja napotyka na trudności związane ze zróżnicowaniem pomiędzy produktami, w szczególności także tymi, które zaliczyć można do tej samej grupy[174].

4.2.3 Kalkulatory ubezpieczeniowe

Ta forma źródła wiedzy charakteryzuje się większą prostotą w stosunku do portali produktowych oferujących pełne kwotowania składki. W praktyce oznaczać to może dwie zasadnicze różnice:

- uproszczenie modelu ze względu na liczbę i zakres zmiennych taryfikacyjnych,
- uproszczenie modelu ze względu na stopień skomplikowania obliczeń.

Pierwsza forma uproszczenia jest zrozumiała ze względu na wygodę potencjalnego nabywcy ubezpieczenia on-line. Podstawowym zadaniem kalkulatora ubezpieczeniowego jest przyciągnięcie klienta do produktu przez orientacyjne wskazanie kosztu jego zakupu.

Druga forma uproszczenia ma charakter bardziej technologiczny. Modyfikacja polegająca na uproszczeniu algorytmu obliczeniowego taryfy ubezpieczeniowej wynikać może z chęci zmniejszenia obciążenia serwera.

[173] ang. „Completely Automated Public Turing test to tell Computers and Humans Apart" - kody wyświetlane w postaci grafik przekształcanych specjalnymi algorytmami w celu zapobieżenia analizy obrazu i odczytaniu ich przez automaty.
[174] Czyli takie produkty, które powinny stanowić bezpośrednio swoje substytuty.

4.3 Rynek ubezpieczeń on-line

W lipcu 2011 roku przeprowadzono badanie rynku ubezpieczeń on-line [Stolarski2012]. W badaniu uwzględniono 61 firm ubezpieczeniowych z siedzibą w Polsce oraz 17 oddziałów przedsiębiorstw zagranicznych. Były to wszystkie oficjalnie prowadzące działalność ubezpieczeniową przedsiębiorstwa w kraju. Wszystkie z przebadanych podmiotów posiadały własne witryny internetowe. Odnotowana liczba witryn oznacza istotny wzrost w stosunku do danych dostarczonych przez podobne wcześniejsze badanie, które przeprowadzone zostało w roku 2006. Wówczas 16,22% podmiotów działu I. oraz 9,09% podmiotów działu II. nie posiadało w ogóle witryny WWW [Kaczała2006].

Taka zmiana odzwierciedla informacje przedstawione w rozdziale 1, dotyczące ogólnego trendu wzrostu wolumenu sprzedaży produktów i usług z wykorzystaniem internetu. Fakt posiadania przez wszystkie podmioty witryny WWW w roku 2011 wskazuje na uwzględnienie przez przedsiębiorstwa ubezpieczeniowe znaczenia funkcji sprzedażowej oraz marketingowej kanału internetowego. Z wyliczeń przedstawionych w cytowanym badaniu z połowy 2011 roku wynika, że w obszarze dystrybucji ubezpieczeń w latach 2006-2011 odbyła się także zasadnicza zmiana. Na początku tego okresu tylko 13,51% podmiotów prowadzących działalność ubezpieczeniową w kraju w ramach II działu ubezpieczeń wykorzystywało kanał internetowy do bezpośredniej sprzedaży przynajmniej jednego produktu. Z kolei na końcu rozpatrywanego w badaniu okresu było to już 57,69% podmiotów z siedzibą w Polsce. Dodatkowo zauważyć należy, że w przeważającej części były to podmioty zorganizowane w formie spółek akcyjnych, które ze względu na konieczność maksymalizacji zysku oraz wartości rynkowej sprawniej dostosowują się do uwarunkowań rynkowych od innych form organizacyjnych.

Jednym z przejawów wykorzystania internetu w procesie produkcji ochrony ubezpieczeniowej jest możliwość zgłoszenia wypadku ubezpieczeniowego oraz prowadzenia działań zmierzających do likwidacji szkód w formie elektronicznej. Według badania z 2011 roku w ramach wykorzystywanego modelu biznesowego 34,62% podmiotów wspierało takie podejście. Przy czym, co można uznać za intrygujące, lista pod-

miotów umożliwiających zgłoszenie szkód on-line różni się od listy dystrybutorów produktów ubezpieczeniowych przez internet[175].

Rynek sprzedaży ubezpieczeń w internecie nie polega wyłącznie na bezpośrednim oferowaniu ochrony przez ubezpieczycieli. Z kanału internetowego aktywnie korzystają także inne podmioty tego rynku, o których wspomniano już w podrozdziale 4.1 (brokerzy, agenci, a także przedstawiciele nurtu bancassurance). Portale porównujące oferty również stają się coraz bardziej popularnym sposobem dostarczania informacji o ofercie produktowej do klientów. Można oczekiwać, że ich znaczenie będzie wzrastać.

4.4 Źródło internetowe a model wyceny

Porównując proponowaną ekstrakcję modelu wyceny ze źródła internetowego z ekstrakcją informacji ze źródeł głębokiego internetu[176], zauważyć można szereg pozornych podobieństw. Najważniejsze różnice między tymi zagadnieniami to:

- w zaproponowanym podejściu ekstrahujemy wiedzę, nie informację; źródłem nie jest baza danych dostępna przez interfejs internetowy, lecz algorytm implementujący model wyceny,
- liczba uzyskiwanych przez zmianę kryteriów kalkulacji wyników jednostkowych w przypadku ekstrakcji modeli wyceny może być zdecydowanie liczniejsza; z drugiej strony: nie wszystkie wyniki jednostkowe przyczyniają się do podniesienia jakości ekstrahowanego modelu,
- w zaproponowanym podejściu ekstrakcja informacji jest częścią składową procesu ekstrakcji wiedzy.

Zadanie ekstrakcji modelu wyceny ubezpieczenia sprowadza się do odtworzenia jak najdokładniejszego przybliżenia algorytmu wyceny składki. Szczegółowe parametry polisy są reprezentowane w postaci wektorów $\boldsymbol{X_i} = [x_{i1} \quad x_{i2} \quad \cdots \quad x_{im}]$. Każdemu takiemu wektorowi odpowiada wielkość składki (wycena) $s_i, 0 < i < n$. Wyceny są pozyskiwane w wyniku symulacji poruszania się (nawigacji) użytkownika po stronie internetowej ubezpieczyciela i wprowadzania zadanych parametrów polisy

[175] tzn. część firm dawała możliwość zgłoszenia szkody, mimo że nie prowadziła sprzedaży produktów ubezpieczeniowych w internecie.
[176] Która omówiona była w podrozdziale 2.2.

$x_{\bullet 1}, \cdots, x_{\bullet m}$. Wartości parametrów wybierane są spośród dostępnych w formularzu internetowym opcji, określanych jako dziedzina parametru $x_{\bullet \mathrm{j}} \in D_j$, $0 < j < m$. Każdy wektor reprezentuje pojedynczy i-ty cykl nawigowania po źródle.

Z każdym pozyskaniem wyceny wiąże się koszt, dlatego też dodatkowym postulatem jest minimalizacja liczby wycen n niezbędnych do odtworzenia modelu. Koszt ten wyrażać się może przez zaangażowanie zasobów, jak też przez potrzebny na wykonanie operacji odpytania czas. Dodatkowym bodźcem związanym z koniecznością minimalizacji liczby otrzymanych wartości funkcji *f* mogą być ograniczenia występujące po stronie źródła.

W zależności od wykorzystanej metody aproksymacji algorytmu (modele regresji, sieci neuronowe, programowanie genetyczne etc.) otrzyma się wynik o różnych postaciach. W szczególnym przypadku będzie to funkcja $f^*()$ wraz z estymatorami jej parametrów $p_1, \cdots, p_k$ (zakładamy, że $k \ll n$). Postulat $n \rightarrow min$ realizowany jest przez redukcję liczebności próbkowanych podzbiorów dziedzin D_m.

W przypadku opisywanego problemu, definicja dziedzin D_m oraz wartości funkcji f^* pochodzić będzie z odpytywania ustalonego oraz reprezentowanego za pomocą odrębnych mechanizmów źródła internetowego. W rezultacie możemy mówić o „zadaniu ekstrakcji modelu wyceny ubezpieczeń ze źródła internetowego". Uogólnieniem problemu jest przeniesienie zadania na wiele heterogenicznych źródeł.

W celu realizacji tak zdefiniowanego zadania potrzebne jest spełnienie szeregu założeń, wśród których najważniejszymi są:

- deterministyczność odtwarzanego modelu,
- jawność istotnych parametrów v_j odtwarzanego modelu oraz ich dziedzin D_j,
- istnienie zdefiniowanej analitycznej postaci funkcji *f*.

Możliwe jest opracowanie metody pozwalającej na wyekstrahowanie modelu ze zróżnicowanych źródeł internetowych. Opracowanie problemu na dostatecznie ogólnym poziomie wymaga ponadto rozwiązania szeregu kwestii, takich jak posiadanie uogólnionego opisu źródła czy założenia dotyczące kształtu samego modelu wyceny. Dodatkowo pomocne jest posiadanie reprezentacji terminologii źródła, np. w postaci ontologii domenowej.

5 Model źródeł internetowych z produktami ubezpieczeniowymi

5.1 Wiedza zakładu ubezpieczeń dot. produktu a wiedza zakodowana w źródle on-line

Jak zauważa autorka artykułu [Sternik2009], „*podstawowym tworzywem produktu ubezpieczeniowego jest wiedza i kreatywność, obie w różnym stopniu zależne od rzetelnej i wszechstronnej informacji*". Firmy niechętnie dzielą się posiadaną wiedzą, określaną także często fachowo jako tzw. know-how. W podrozdziale 3.3 przedstawiono zestawienie informacji o źródłach zasilających procesy ubezpieczycieli związane z tworzeniem oraz aktualizacją produktów ubezpieczeniowych pod kątem taryfikacji i składki. Źródłami takimi są źródła wewnętrzne, tj. systemy bazodanowe, hurtownie danych oraz dedykowane systemy business intelligence. Szczególnie w przypadku nowych produktów źródłami takimi mogą być źródła zewnętrzne, tj. publicznie dostępne lub komercyjnie pozyskane dane.

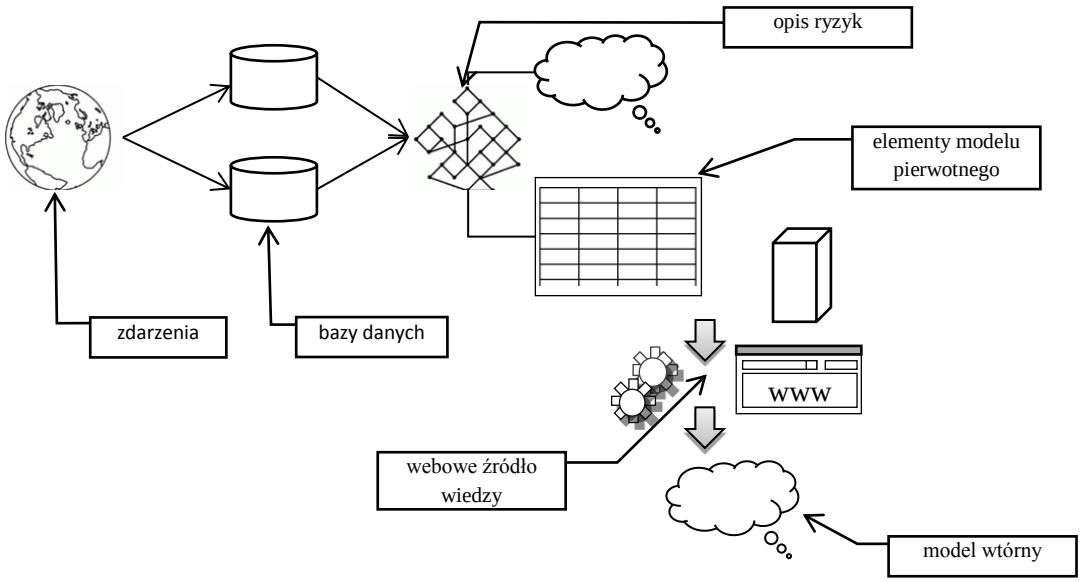
Niezależnie od pochodzenia, zakumulowane dane mają charakter historyczny i opisują rzeczywistość związaną z określoną grupą przedmiotów ubezpieczenia. Na podstawie takich danych tworzona jest dokumentacja dla opracowywanego produktu oraz wyznaczana jest zależność pomiędzy ryzykiem a ceną nabywanego produktu. Zależność, o której tu mowa, może mieć bardzo zróżnicowaną postać. Szczegółowe metody służące do wyznaczania tej zależności zaprezentowano w podrozdziałach 3.2.5-3.2.8. W rozumieniu pracy takie przekształcenie danych i informacji o historii rzeczywistych obiektów w abstrakcyjną wiedzę polega na stworzeniu modeli opisujących subiektywne ryzyko zakładu ubezpieczeń związanego z przyjęciem podobnych przedmiotów do ubezpieczenia, co przy zestawieniu z kosztami funkcjonowania zakładu jako podmiotu gospodarczego przekłada się na **model pierwotny - taryfikacji (model wyceny składki)**[177].

Gruntowna dyskusja funkcjonowania kanałów dystrybucji w ubezpieczeniach przeprowadzona została w podrozdziale 3.1.2-3.1.3. Każdy kanał dystrybucji posiada

[177] W podrozdziale 3.2.5 wspomniane jest, iż w gruncie rzeczy przy projektowaniu produktu ubezpieczeniowego istnieją zazwyczaj dwa algorytmy – jeden przeznaczony do *underwritingu*, czyli dopuszczenia danego ubezpieczenia oraz drugi do faktycznej kalkulacji składek. W przypadku źródeł webowych oba algorytmy sprowadzone są do jednego procesu.

dostęp do modelu taryfikacji, przy czym implementacje modelu w postaci odpowiedniego narzędzia wyposażonego w określony algorytm oraz niezbędne informacje mogą podlegać pewnym modyfikacjom. Modyfikacje te wynikają ze specyfiki kanałów oraz aspektów marketingowych przedstawionych w podrozdziale 3.1 i następnych. Algorytm wyceny składki może pochodzić bezpośrednio z systemu produktowego lub stanowić kopię pochodzącą z systemu do ofertowania [Sternik2009]. Częściej jednak ze względu na uproszczony charakter powinien być osobną wersją modelu dającą mimo to zbliżone wyniki.

Model taki ma w przebiegu zdefiniowanego w pracy zadania oraz metody ekstrakcji charakter wzorcowy. Oznacza to, że na podstawie próby uzyskanej z generowanych przez taki model wyników próbujemy zbudować nowy model, który nazywać będziemy **modelem wtórnym lub wyekstrahowanym**. Całość przedstawionych powyżej rozważań oraz definicji zobrazowana została na rysunku 9.



Rysunek 9. Elementy modelu pierwotnego wyceny składki a model wtórny
Źródło: opracowanie własne

Zestawiając modele pierwotny oraz wtórny zastanowić należy się nad spodziewanym podobieństwem oraz ewentualnymi różnicami. Rezultat takiej analizy zademonstrowano w tabeli 4.

Tabela 4. Zestawienie cech modeli: pierwotnego oraz wtórnego
Źródło: opracowanie własne

Model	Pierwotny	Wtórny
Postać	nie do określenia	zróżnicowana
Wyniki	mierzalne	zbliżone lub takie same
Zmienne	podane explicite	takie same
Błąd	brak	możliwy do oszacowania

5.2 Założenia wstępne i ograniczenia

Przeprowadzanym badaniom prezentowanym w niniejszej pracy oraz opracowywanej metodzie towarzyszyły następujące założenia:

- przedmiotem badań są źródła należące do dobrze zdefiniowanego zbioru źródeł,
- każde z badanych źródeł charakteryzuje się istnieniem związanego z nim poznawalnego algorytmu – model ubezpieczyciela oraz działanie źródła wiedzy są w pełni deterministyczne,
- zwiększenie liczby danych może prowadzić do dokładniejszego odtworzenia modelu,
- istotne informacje dotyczące klienta zasilają model ubezpieczyciela wyłącznie w sposób jawny w wyniku kolejnych kroków procesu kalkulacji składki,
- zakresem badań objęte są wyłącznie modele kalkulacji składki ubezpieczenia o określonej budowie i strukturze modelu wyceny,
- występuje powiązanie wyników modelu ubezpieczyciela z rzeczywistym ryzykiem.

Poniżej odnosimy się do poszczególnych założeń.

Zbiór badanych źródeł

Do badania wybrano zróżnicowany zestaw źródeł spośród rodzajów omówionych w podrozdziale 4.2. Badanie obejmuje wyłącznie polski internet. Co prawda wybrane źródła zagraniczne (witryny pochodzących z określonych krajów lub obszarów języ-

kowych) były rozważane jako kandydaci do rozszerzonej wersji badania, jednak ostatecznie podjęto decyzję o ograniczeniu zakresu badania[178].

Źródła dobrane są w oparciu o szereg wyszczególnionych kryteriów szczegółowo przedstawionych w podrozdziale 4.1. W ramach doboru uwzględniono także czynniki technologiczne z podziałem na typy technik implementacji witryn internetowych prezentowanych w podrozdziale 2.1. Dodatkowo dobór uwzględnia bariery oraz ograniczenia mające wpływ na dostępność oraz skuteczność automatyzacji procesu ekstrakcji. W tym ostatnim wypadku ocena dopuszczalności ma charakter akceptacji poprzez zastosowanie metody eksperckiej.

Poznawalność algorytmu w źródle

Drugie w kolejności założenie odnosi się bezpośrednio do racjonalności gospodarczej podmiotów działających na rynku. Instytucje ubezpieczeniowe, tworząc produkt, są zobligowane do wykorzystania w tym celu metod aktuarialnych. Obszerne zestawienie stanu wiedzy dotyczącego tychże metod podano w podrozdziałach 3.2.2-3.2.8. Algorytmy oraz inne artefakty składające się na model taryfikacji są dokumentowane oraz archiwizowane w systemach informacyjnych przedsiębiorstwa [Sternik2009], co zasadniczo nakłada wymóg nadawania modelom semantycznej postaci.

Spośród wspomnianych metod aktuarialnych obecnie największym znaczeniem cieszą się metody wielowymiarowe opisane w podrozdziale 3.2.7 ze szczególnym uwzględnieniem uogólnionych modeli liniowych. Rezultaty zastosowania tej lub opartych na niej metod gwarantują stabilną i deterministyczną postać modelu kalkulacji składki wyrażoną za pomocą formuł matematycznych.

Można rozważać teoretycznie systemy produktowe dostarczające modele wyceny składki aktualizowane w czasie rzeczywistym. W praktyce jednak utrzymanie takiego rozwiązania powodowałoby zapewne znacznie więcej kosztów dla ubezpieczyciela niż korzyści. Taryfy oraz modele z nimi związane są zazwyczaj aktualizowane w określonych odstępach czasu (przynajmniej rocznie ze względu na konieczność uwzględnienia efektów inflacyjnych).

[178] Na powody takiej decyzji składał się fakt ograniczeń czasowych, możliwości technicznych oraz dostatecznie zadowalające rezultaty otrzymane z badania samych źródeł krajowych. Dodatkowo takie ograniczenie pozwoliło precyzyjnie określić granice prowadzonych prac badawczych. W przypadku podjęcia decyzji przeciwnej wyznaczenie takiej granicy byłoby problematyczne.

Zależność dokładności modelu od liczby danych

Zastosowane metody rekonstrukcji modelu wtórnego na podstawie danych pochodzących z modelu pierwotnego mają charakter metod ilościowych lub wywodzą się z dziedziny systemów ewoluujących i uczących się (sztuczna inteligencja). W obydwu przypadkach liczba danych wejściowych przy tworzeniu modelu ma kluczowy[179] wpływ na wyniki w postaci dokładności otrzymanego na wyjściu modelu wtórnego. Metoda regresji, będąca przedstawicielem metod statystycznych wykorzystanych w badaniu, scharakteryzowana została w podrozdziale 2.4.1. Polega ona na szacowaniu parametrów w modelu o określonym schemacie. Zakłada się, że estymatory tych parametrów są zgodne. W takim przypadku uczynione w badaniu założenie o zależności dokładności modelu należy uznać za zasadne. Zależność, o której mowa, nie jest tak oczywista w przypadku drugiej grupy metod wykorzystujących techniki sztucznej inteligencji. Metody te są omówione w podrozdziale 2.4.2-2.4.3. W ramach wspomnianego omówienia tematyki tychże narzędzi odniesiono się także do zagadnienia wielkości próby.

Jawność informacji zasilających źródło

Współczesna technologia komunikacji za pomocą protokołu HTTP(S), czyli wymiany informacji ze źródłami webowymi, potencjalnie daje szereg możliwości do zbierania dodatkowej informacji o użytkownikach biorących udział w procesie komunikowania. Do powszechnych sposobów pozyskiwania informacji nie wprost przez dostawców treści w kanale internetowym zaliczyć można przede wszystkim analizę adresów IP, a w szczególności tzw. geolokalizację. Geolokalizacja jest mechanizmem dającym możliwość poznania – często z dość dużą dokładnością – lokalizacji użytkownika na podstawie adresów sieciowych. Wykorzystywanie narzędzi geolokalizacyjnych ma powszechny charakter. Starsze wersje przeglądarek dawały możliwość konstrukcji profilu użytkownika w oparciu o dane z historii odwiedzanych witryn czy zapisane w plikach *cookies*. Tego rodzaju metody zostały jednak uznane za szkodliwe, a praktyki zmierzające do pozyskania informacji bez wiedzy użytkownika stanowią naruszenie prywatności. Stąd nowsze wersje przeglądarek zapobiegają tego rodzaju możliwościom lub też obsługują protokoły komunikacji umożliwiające deklarację za-

[179] Pomijając aspekt jakości.

kresu zbieranych przez serwery danych o użytkowniku. To ostatnie rozwiązanie możliwe jest np. przez wykorzystanie standardu P3P[180]. W konsekwencji licznych przypadków naruszeń prywatności przez witryny webowe istotna liczba prawodawstw wprowadziła regulacje dotyczące tego rodzaju działań. Przykładowo w Europie działania polegające na próbach wykrycia historii przeglądanych witryn lub zbieraniu innych poufnych informacji bez wiedzy użytkownika uznawane są za nielegalne[181].

Mimo to firmy ubezpieczeniowe mogą mieć możliwość pozyskania pewnych dodatkowych informacji związanych z profilem użytkownika lub nawet szczegółowych danych związanych z klientem. Sytuacja taka może mieć miejsce np. w przypadku kanału bankassurance on-line, gdy potencjalny klient ma już założone konto bankowe, a następnie (zalogowany lub nawet nie – o ile transakcje odbywają się w tej samej domenie) dokonuje zakupu lub tylko dokonuje przeglądu ofert. Innym przypadkiem może być wykorzystanie koincydencji faktów na podstawie posiadanych lub dostarczonych informacji przez klienta[182].

W przeprowadzonym badaniu zakładamy, że takie przypadki nie mają miejsca, a jeżeli mają, to nie uwzględniamy ich efektów. Powodem, dla którego przyjęte zostało takie założenie, jest niemożliwość stwierdzenia ani wykrycia tego rodzaju ewentualnych działań. Ostatecznie należy stwierdzić, że nawet jeżeli są one rzeczywiście stosowane, to przypadki ich występowanie mają charakter marginalny.

Właściwa postać i struktura algorytmu w źródle

To założenie oznacza, że algorytm kalkulacji ma postać wzoru matematycznego operującego na znanej liczbie zmiennych (zmiennych taryfikacyjnych). Typ każdej zmiennej taryfikacyjnej należy do zamkniętego zbioru typów zmiennych, które zestawione zostały w tabeli 5.

[180] ang. The Platform for Privacy Preferences. Dokumentacja znajduje się na stronie: http://www.w3.org/TR/P3P/, odczytano 30-05-2015 r.

[181] Por. dyrektywa 2000/31/WE Parlamentu Europejskiego i Rady z dnia 8 czerwca 2000 r. o handlu elektronicznym.

[182] Jeszcze inny aspekt omawianego problemu zauważalny jest w kontekście intensywnego rozwoju Web 2.0, a w szczególności sieci społecznych. Nie od dziś wiadomo, że informacje zamieszczane przez samych użytkowników na temat swój lub osób powiązanych w ramach tzw. profilu użytkownika, na portalach typu Facebook czy Nasza Klasa, mogą prowadzić do sytuacji zagrożenia prywatności. Autorzy tekstu [Abramowicz2011] rozważali już możliwość wyszukiwania danych – aczkolwiek o innym charakterze – w internecie w celu wsparcia niektórych procesów w branży ubezpieczeń. W świetle jednak doniesień prasowych, http://pej.cz/WIDEO-Alior-Bank-wie-o-swoich-klientach-wszystko-i-z-checia-to-sprzeda-a5618 okazuje się, że wizja ta nie tylko nie jest daleka od spełnienia, ale przyjęła wręcz nieco groteskowy charakter. Odczytano 02-03-2013 r.

Tabela 5. Zakładane typy zmiennych taryfikacyjnych
Źródło: opracowanie własne

Typ zmiennej taryfikacyjnej	Podtyp	Uwagi
Logiczny	Dwustanowy	Np.: tak, nie; prawda, fałsz.
	Trójstanowy	Trzeci stan nieustalony.
Liczbowy	Dyskretny	
	Ciągły	
Enumeratywny		Zamknięty zbiór wartości.
Tekstowy		
Czasowy		Daty w różnorodnych formatach.
Specjalny		Zdefiniowany w ontologii.

Oprócz tego algorytm kalkulacji może operować na zamkniętym zbiorze danych sprowadzalnych do postaci tablicy. Dane takie mogą reprezentować mnożniki, korekty w postaci zwiększeń lub umniejszeń etc.

Powiązanie wyników taryfikacji oraz ryzyka

To założenie oznacza, że oczekiwana jest zgodność modelu pierwotnego ze statystykami uzyskanymi przez towarzystwo ubezpieczeniowe w trakcie procesów związanych z prowadzeniem biznesu. Oznacza to także, że wraz ze zmianami parametrów ryzyk w czasie odpowiadający temu model będzie także podlegał modyfikacjom. Zgodnie z opisem przedstawionym w podrozdziale 3.1.2 oraz 3.2.2, model taryfikacji produktu ubezpieczeniowego odpowiada faktycznym parametrom ryzyk związanych z tymże produktem. Niemniej istnieje szereg czynników, wymienianych m.in. w podrozdziale 3.2.6, które zaburzają rzeczywisty obraz ryzyk reprezentowanych przez model.

Weryfikacja prawdziwości założenia w przeprowadzonym badaniu może częściowo odbyć się przez analizę wiedzy zewnętrznej w stosunku do modelu, pochodzącej z innych źródeł danych dotyczących przedmiotu ubezpieczenia lub konkurencyjnych modeli[183].

[183] Część wyników zaprezentowanych w rozdziale 7 pośrednio potwierdza słuszność tego założenia. Chodzi przede wszystkim o fakt poprawy wyników modeli wzbogaconych o dane zewnętrzne, tak jak to tutaj zaproponowano.

5.3 Metoda modelowania oraz decyzje dotyczące kształtu modelu

W celu zbudowania skutecznego narzędzia ekstrakcji danych z portalu ubezpieczeniowego konieczne było stworzenie zaawansowanego modelu takiego portalu w celu umożliwienia automatycznej nawigacji po nim. Prace koncepcyjne przy modelowaniu polegały na dwóch rodzajach działań. Pierwszym działaniem była analiza innych, podobnych modeli, tworzonych na potrzeby systemów ekstrakcji informacji. Prace opisujące takie systemy oraz sposoby rozwiązywania określonych problemów z nimi związanych szczegółowo przedstawiono w podrozdziale 2.2.1-2.2.2. Drugim rodzajem działań były eksperymenty związane z implementacją autorskiego rozwiązania [Stolarski2012]. Rezultatem obydwu postępowań było podjęcie szeregu decyzji projektowych dotyczących tworzonego narzędzia wspierającego realizację zadania, jak również mających bezpośredni wpływ na samą metodę. Zestawienie tychże decyzji zawiera tabela 6.

Model źródła portalu sprzedaży ubezpieczeń zastosowany w naszym prototypie badawczym posiada dwa wymiary. Pierwszy wymiar dotyczy opisu źródła w zakresie nawigacji pomiędzy lokalizacjami w ramach portalu. Drugi wymiar związany jest z opisem nawigacji w ramach lokalizacji. Może on dotyczyć zarówno działań – akcji symulujących interakcję użytkownika, jak również odzwierciedlać dynamiczną naturę treści w danej lokalizacji.

Do opisu modelu zdecydowano się wykorzystać istniejące i sprawdzone standardy – przede wszystkim formalizm XML. Powodem jest elastyczność oraz możliwość łatwego przechowywania zawartości obiektów programowych (serializacja). Dodatkowo zaletą przechowywania informacji w języku XML jest możliwość dostępu do informacji także przez człowieka[184], a także możliwość bezproblemowego łączenia z innymi, użytecznymi z punktu widzenia projektu i zgodnymi formalizmami: opisu ontologii (OWL) oraz stron WWW (XHTML, XPath, XSLT), czy wreszcie „wstrzykiwanie”[185] skryptów.

Nawigacja pomiędzy lokalizacjami reprezentowana jest za pomocą typowanego, skierowanego grafu, w którym poszczególne typy wierzchołków odpowiadają artefak-

[184] Istotny aspekt na etapie wykonywania eksperymentów oraz rozwoju systemu.
[185] ang. injection. Technika kojarzenia treści zapisanych w różnych formalizmach, w szczególności treści luźno powiązanych.

tom związanym z nawigacją po źródle. W tabeli 7 zestawiono typy składowych elementów nawigacji.

Tabela 6. Decyzje projektowe dotyczące zasad tworzonia prototypu rozwiązania
Źródło: opracowanie własne

Decyzja	Zakres	Podobieństwo	Alternatywa
Dwuetapowość procesu	Faza przygotowania, faza wykonawcza		Jednofazowy przebieg – utrudnione tworzenie opisu źródeł. Podział na trzy fazy z oddzielną fazą generowania modeli wtórnych
Reprezentacja przy pomocy grafu	Etapy budowy grafu, uszczegółowienia grafu, ekstrakcja danych	[Kaczmarek2006]	Wykorzystanie formalizmów przepływu pracy lub ontologizacja opisu
Opis semantyczny	Etapy adnotacji pojęciami, decyzji optymalizacyjnych, ekstrakcji danych	-	
Wsparcie dla tworzenia opisu źródeł	Faza przygotowawcza	[Baumgartner2001]	Brak fazy przygotowawczej
Koncentracja na ekstrakcji strukturalnej	Etap ekstrakcji danych	[Flejter2011]	Implementacja mechanizmów ekstrakcji informacji z tekstu [Węcel2011]
Podejście półautomatyczne	Faza przygotowawcza	[Arasu2005]	Algorytmy uczenia maszynowego [Kushmerick1997]

Wybrane typy zostaną omówione szczegółowo w podrozdziale 6.2. Architektura systemu obrazująca powiązania pomiędzy powyżej przedstawionymi typami została przedstawiona na diagramach zawartych w aneksie C („Metoda ekstrakcji – schematy UML").

5.4 Struktury danych

W niniejszym podrozdziale prezentujemy struktury danych wykorzystywane przez prototyp narzędzia realizującego metodę ekstrakcji. Opisane zostaną następujące struktury, mające jednocześnie kluczowe znaczenie dla realizacji funkcjonalności narzędzia: właściwości, warunkowe zbiory wartości, wzorce nawigacyjne, konkretyzacje wzorców.

Tabela 7. Elementy składowe grafu nawigacji
Źródło: opracowanie własne

Typ elementu	Funkcja
Podstrona (AutomationSiteNode)	Wywołanie żądania dostarczenia treści dokumentu o określonym URL za pomocą metody GET
Miernik czasu (AbstractTimer)	Wierzchołek realizujący logikę nawigacji polegającą na odczekaniu określonego przedziału czasowego. Przedział zależy od implementacji podtypu oraz kontekstu użycia. Podtypy: LoadAjaxTimerNode, RandomTimerNode, ReloadTimerNode, StaticTimerNode, UserTimerNode.
Serwer proxy (AbstractProxy)	Rozpoczęcie nawigacji za pośrednictwem zdalnej jednostki pośredniczącej. Serwer Proxy dobierany jest na podstawie wybranej strategii. Podtypy: AbstractMultiProxy, MultiConnectionProxy, WebServiceFedMultiProxy.
Węzeł warunkowy (PropertyConditionalNode)	Realizuje rozgałęzienia w grafie nawigacji na podstawie zadanych warunków.
Warunek (PropertyCondition)	Reprezentuje warunek w węźle warunkowym.
Wartość warunkowa (ConditionalValue)	Element warunku w węźle warunkowym.
Węzeł automatyzujący (AutomationSiteNodePattern)	Wzorzec nawigacji definiujący listę zautomatyzowanych operacji.
Pojedyncze zadanie automatyzacji (AutomationTask)	Jednostka zadania automatyzacji. Bardziej szczegółowy opis w podrozdziale 6.2.
Węzeł wykonywujący skrypt (JavaScriptExecutor)	Umożliwia zdefiniowanie dowolnego kodu skryptu, który zostanie wykonany w środowisku strony.
Element startowy (Source)	Początkowy element rozpoczynający graf nawigacji.
Element końcowy (Terminator)	Element kończący graf nawigacji.
Stan automatu nawigującego (AbstractSiteNodeState)	Struktura przechowująca informacje o aktualnym stanie automatu nawigującego.
Abstrakcyjny wzorzec nawigacji (AbstractSiteNodePattern)	Wierzchołek reprezentujący wzorzec nawigacji.
Abstrakcyjna konkretyzacja wzorca nawigacji (AbstractSiteNode)	Konkretyzacja wzorca nawigacji.
Warunek logiczny (AbstractConditionalNode)	Wierzchołek w grafie nawigacji realizujący określoną logikę – warunkowe podążenie określonymi krawędziami
Abstrakcyjny wierzchołek ekstrahujący (AbstractSiteExtractor)	Pozwala wyodrębnić treści ze strony na podstawie określonych reguł. Podtypy: AbstractSiteRegexExtractor, AbstractSiteXPathExtractor, AbstractSiteXsltExtractor.

Właściwości

Właściwość (property) jest reprezentacją i odpowiednikiem zmiennej taryfikacyjnej w modelach kalkulacji składki (zarówno w modelu pierwotnym, jak i wtórnym). Pojęcie zmiennej taryfikacyjnej wprowadzane jest w podrozdziale 3.2.6 poświęconemu podstawowym technikom aktuarialnym.

Tabela 8. Rodzaje i opis właściwości
Źródło: opracowanie własne

Rodzaj właściwości	Typ zmiennej	Uwagi
AbstractIterable	dowolna	Generalizacja
ConditionalTextualIterable	tekstowa	Z wartościami warunkowymi
NumericIterable	numeryczna	Implementująca operator iteracji
NumericRandomGenerated	numeryczna	Losowo generowana przy każdej iteracji
TextualIterable	tekstowa	Implementująca operator iteracji
TextualRandomGenerated	tekstowa	Losowo generowana przy każdej iteracji
TextualRandomValued	tekstowa	Generowana losowo dla procesu
TextualOntologicalIterable	specjalizowana	Implementująca operator iteracji
TextualOntologicalRandomGenerated	specjalizowana	Losowo generowana przy każdej iteracji
DateTimeIterable	czasowa	Generuje daty w określonym formacie i z określonym odstępem
PropertyFormatter	dowolna	Właściwość „techniczna”. Umożliwia dokonywanie operacji tekstowych i użycia wyrażeń regularnych na wartościach innych właściwości

Jak zauważono przy okazji analizy tychże technik, w modelach taryfikacyjnych wykorzystuje się różne typy zmiennych dla celów modelowania. Stąd proponowana reprezentacja także odzwierciedla takie potencjalne zróżnicowanie. Tabela 8 przedstawia rodzaje właściwości oraz ich odpowiedniki w modelu pierwotnym mające zastosowanie w prototypowym narzędziu.

Warunkowe zbiory wartości

Zaproponowany w pracy model źródła obejmuje dwa wymiary. W poprzednim podrozdziale przedstawiono pierwszy wymiar, tj. nawigowanie pomiędzy lokalizacjami witryny WWW. Dla opisu drugiego wymiaru posłużono się specyficzną strukturą, którą określić można jako drzewo z warunkami. Jest to drzewo, w którym krawędzie oznaczone są prostymi formułami warunkowymi, których operandami są wartości

z wierzchołków drzewa, przy czym brak spełnienia warunku zapisanego przy danej krawędzi powoduje ignorowanie poddrzew wywodzących się z tejże krawędzi.

Wzorce nawigacyjne

Wzorce nawigacyjne są niezależnymi elementami składającymi się na proces nawigacji po źródle webowym. W swojej istocie są one najbliższe terminowi wzorca akcji webowej (Web Action Template) wprowadzonemu w [Flejter2011]. Jednak występuje tutaj szereg różnic. Różnice te wynikają z bardziej realistycznych założeń architektonicznych dotyczących opisywanego rozwiązania. Przy czym są one mniej ogólne w swoim rozwiązaniu, ze względu na zakres projektu oraz specyfikę zadania[186].

Strukturę wzorca nawigacyjnego zapisać można schematycznie w następujący sposób:

$$Wn = \{u(sm), S, En, P, g = (g_i, s_i), p = (p_j, s_j)\}, \quad (14)$$

gdzie:

- *u* jest poprawnym szablonem adresu URI lub ciągiem pustym,
- *S* jest strukturą reprezentującą stan układu klient-serwer,
- *En* jest zbiorem reguł ekstrakcji,
- *P* jest elementem zbioru reprezentującego serwery proxy lub elementem pustym,
- *g* jest ciągiem par parametrów oraz symboli właściwości przeznaczonym dla metody komunikacji GET,
- *p* jest analogicznym ciągiem dla metody komunikacji POST.

Struktura reprezentująca stan układu klient-serwer ma postać:

$$S = \{(s_1, w_1) \ldots (s_x, w_x), c = (c_v, s_v)\}, \quad (15)$$

gdzie:

- (s_n, w_n) to ciąg par symboli właściwości oraz nadanej wartości dla odpowiadającej właściwości,
- *c* jest zbiorem informacji z ciasteczkami.

[186] W przypadku podjęcia próby zbudowania silnika o bardzo dużej – jeżeli nie granicznie możliwej – ogólności, nawigacja po źródle mogłaby być zrealizowana za pomocą silnika przetwarzającego semantykę któregoś z formalizmów opisu przepływów pracy (XPDL, jBPM). Języki te zapewniają możliwość realizacji praktycznie dowolnego przebiegu oraz interakcji. W trakcie realizacji opisywanego badania taka możliwość była brana pod uwagę. Nie miałoby to jednak istotnego wpływu na oczekiwane rezultaty badania.

W przyjętym rozwiązaniu abstrahujemy od dualizmu stanowości klienta oraz serwera. W zamian zakładamy, że przejście pomiędzy stanami nawigacji zależne jest tylko od stanu serwera oraz zadanego zapytania. Taki model nawigacji jest równoważny modelowi z rozróżnieniem stanów na stronę kliencką oraz serwerową, ale jego zaletą jest większa prostota (mniej elementów) oraz większa zgodność z rzeczywistym funkcjonowaniem komunikacji pomiędzy przeglądarką internetową a serwerem[187].

Konkretyzacje wzorców

W fazie wykonywania nawigacji po źródle webowym wzorce nawigacyjne przeobrażane są w konkretyzacje wzorców. Konkretyzacja wzorca jest strukturą analogiczną do samego wzorca nawigacyjnego, z którego zostaje wyprowadzona. Zawiera ona jednak nie tyle potencjalne informacje o jednostce komunikacji z serwerem webowym, co gotowy zbiór informacji niezbędny do wykonania i obsługi zapytania do serwera. Dane opisujące konkretyzację wzorca mają następującą strukturę:

$$\Lambda(Wn) = \{\lambda(u(sm), S, P), c = (c_v, w_v), En, P, g = (g_i, w_k), p = (p_j, w_l)\}, \quad (16)$$

gdzie:

- w przypadku ciągów *c*, *p*, *g* dla par podstawiane są wartości właściwości odpowiadających poszczególnym symbolom tych właściwości,
- λ jest przekształceniem szablonów adresów URI do przestrzeni adresów serwera proxy, jednocześnie zastępując wartościami właściwości występujących w szablonie adresu symboli właściwości.

[187] Wprowadzenie rozróżnienia pomiędzy stanami klienta a stanami serwera miałoby praktyczny sens wówczas, gdyby poszczególne przejścia nawigacji stanów były równorzędne, tj. możliwe byłoby przejścia z dowolnej kombinacji stanu klienta oraz serwera do innej kombinacji. W praktyce nie obserwuje się takiej możliwości w interakcji przeglądarki z serwerami. Decydujące znaczenie w dostępie do wyświetlania określonych treści (dokumentów) ma stan serwera. W rezultacie stan klienta praktycznie zawsze jest powiązany ze stanem serwera a jedyny istotny wpływ na kolejne przejście mają informacje zawarte w zapytaniu.

6 Metoda ekstrakcji modeli wyceny składki ze źródeł internetowych

6.1 Dobór źródeł wyceny produktu ubezpieczeniowego

Potencjalne źródła webowe użyteczne dla celów badanej metody zostały sklasyfikowane i podzielone według szeregu kryteriów. Kryteria te w ogólności mają znaczenie techniczne oraz biznesowe[188]. W podrozdziale 2.1 wskazano na typowe rozwiązania stosowane przy tworzeniu witryn internetowych z punktu widzenia wykorzystanych przy ich tworzeniu lub w ich działaniu technologii. W podrozdziale 4.1 zaprezentowano szerokie spektrum typologii ubezpieczeniowych źródeł informacji w internecie według kryteriów modeli biznesowych, w których źródła te uczestniczą oraz kompletności procesów dystrybucji. Z kolei w podrozdziale 4.2 wskazano na rozróżnienie źródeł wiedzy ze względu na stopień zaawansowania i liczbę ofert algorytmów taryfikujących. Wreszcie w podrozdziale 5.2 wspomniano kwestię zróżnicowania portali ze względu na kryterium geograficzne.

Zestawienie wszystkich tych kryteriów tworzy przestrzeń decyzyjną pozwalającą na zidentyfikowanie witryn, które mogą[189] stanowić źródła wiedzy wykorzystywane przez proponowaną metodę. W celu uzyskania jak najlepszej jakości i wiarygodności badania przedstawione poniżej wytyczne odnośnie wyboru źródeł zostaną następnie praktycznie wykorzystane. Jednocześnie w ramach wyszczególnionych kryteriów podjęte zostaną kroki zmierzające do zdywersyfikowania badanych witryn w charakterze źródeł wiedzy. Zakłada się, że taka dywersyfikacja ma pozytywny wpływ na praktyczne wykazanie istotnego poziomu ogólności metody.

Dobór źródeł i zasady ich selekcji pod względem technologicznym

Należy zauważyć, odnosząc się do wskazanych modeli źródeł internetowych, że duża część serwisów ubezpieczycieli oferujących ubezpieczenia on-line posiada cechy serwisów z zaawansowanym GUI. Praktycznie wszystkie tego typu źródła korzystają także z podstawowych zabezpieczeń oferowanych przez szyfrowane połączenie za pomocą protokołu HTTPS. W szczególności portale porównujące oferty mają

[188] Wszystkie te zestawienia razem tworzą siatkę (kostkę), na podstawie której realizowany będzie dobór instancji przedmiotu do przeprowadzenia badań.

[189] W dalszej części dyskutujemy ograniczenia, które są zastosowane do przeprowadzonego badania. Ze względu na dbałość o jakość badania i wyników przyjęte ograniczenia mają charakter zaostrzony w stosunku do normalnych warunków użycia metody.

szereg cech wspólnych ze źródłami głębokiego internetu[190]. Rozpatrując kryterium technologiczne, w badaniu nie zakłada się wprowadzania szczególnych ograniczeń. Jeżeli ograniczenia takie występują, to ze względu na zaobserwowane na podstawie przeglądu szeregu przykładów zależności, jak na przykład brak konieczności obsługi mechanizmów personalizacji. Wśród wszystkich wiodących witryn towarzystw ubezpieczeniowych nie natrafiono bowiem na żadne rozwiązanie tego typu[191]. Nie planuje się także obsługi rozwiązań stworzonych na bazie szkieletów lub oprogramowania uruchamianego po stronie klienta o charakterze multimedialnym. Przede wszystkim chodzi tutaj o formularze działające w technologii Adobe Flash oraz Microsoft Silverlight czy apletów Java. Przypadki zastosowania pierwszego z wymienionych rozwiązań można sporadycznie spotkać w praktyce witryn ubezpieczeniowych[192].

Dobór według kryterium modelu biznesowego

Omawiając podział portali oferujących ubezpieczenia według kryterium modelu biznesowego zaznaczono, że tylko część portali charakteryzuje się możliwością dostępu na zasadzie powszechności. Opisywana metoda koncentruje się na witrynach, które są dostępne w ten sposób. Oznacza to, że zasadniczo z badania wykluczone są witryny intranetowe (B2E) oraz przeznaczone dla biznesu (B2B). W pierwszym przypadku źródło nie ma charakteru powszechnie osiągalnego, a co za tym idzie, dostęp do niego jest zazwyczaj istotnie utrudniony lub całkowicie niemożliwy. Wyklucza to także możliwość obiektywnej oceny wyników ewentualnych badań czy odtworzonych modeli. W drugim przypadku powód decyzji o wykluczeniu spowodowany jest rzadkim wykorzystywaniem medium internetowego (ubezpieczenia direct) w sprzedaży produktów ubezpieczeniowych w segmencie biznesowym oraz korporacyjnym. Na podstawie analizy licznych przypadków ofert firm polskich oraz zagranicznych w zasadzie nie stwierdzono takich przypadków.

W segmencie sprzedaży masowej, skierowanej do konsumenta indywidualnego, oferta produktowa jest za to bardzo szeroka. W ramach modelu biznesowego B2C w podrozdziale 4.1 wyróżniono 5 typów witryn obsługiwanych bezpośrednio lub przez pośredników. Wskazano także, że z punktu widzenia wiarygodności i jakości możli-

[190] Listy wyników, stronicowanie (paginacja) etc.
[191] Pominięto tutaj kwestie portali klienckich. Nie służą one jednak bezpośrednio do ofertowania produktów.
[192] np., www.travelguide.com, odczytano 03-02-2012 r.

wych do otrzymania modeli wyceny najbardziej istotne są witryny prowadzone przez zakłady ubezpieczeń. Niektóre inne typy stron z zasady nie stanowią źródła wiedzy, ponieważ nie zawierają szczegółowych informacji o ofercie.

W rezultacie strategia wyboru witryn dopuszczonych do badania jest następująca: podstawowym przedmiotem badań są witryny dystrybucji bezpośredniej; nie będą natomiast wykorzystane witryny pośredników niezależnych; nie będą także przedmiotem bezpośredniego zainteresowania wortale tematyczne. Jedną z grup podmiotów potencjalnie mogących wykorzystać prezentowaną metodę mogą być właśnie podmioty prowadzące witryny pośredników niezależnych, jak i wortale tematyczne[193]. Stąd w ramach strategii wyboru witryn podjęto decyzję o nie włączaniu przypadków takich stron do zbioru źródeł stanowiących przedmiot badania.

Dobór ze względu na stopień zaawansowania i liczbę stosowanych w ofercie algorytmów taryfikujących

Rozpatrując kryterium zaawansowania i liczby stosowanych w ofercie algorytmów taryfikujących, zakłada się priorytetowe traktowanie witryn będących źródłem pojedynczego algorytmu o oczekiwanej najlepszej jakości generowanych wyników. Jak zauważono przy opisie tego podziału, takie warunki propagacji wiedzy zapewniają przede wszystkim strony będące własnością towarzystw ubezpieczeniowych, które za ich pośrednictwem oferują swoje produkty on-line. Spośród wymienionych w podrozdziale 4.1 w ramach odnośnego kryterium rodzajów stopnia zautomatyzowania procesu sprzedażowego za najbardziej właściwe źródła dla badanej metody uznano witryny internetowe wspierające pełen proces zakupu ubezpieczenia. Do tego podzbioru witryn przynależą także rozróżnione witryny, które dodatkowo pozwalają na obsługę posprzedażową klientów. Z punktu widzenia badanej metody ta ostatnia funkcjonalność jest nieistotna.

Dobór według kryterium geograficznego

Założenie ogólności metody ekstrakcji wiedzy rozumieć należy jako możliwość jej zastosowanie na względnie szerokim zbiorze przedmiotów badania, idealnie o zdywersyfikowanych cechach. Dywersyfikacja cech przedmiotów badanych wynika

[193] Kwestia zastosowań metody będzie przedmiotem dyskusji po prezentacji wyników w podrozdziale 7.6 oraz rozdziale 8.

po pierwsze ze zróżnicowania podmiotów prowadzących witryny ubezpieczeniowe. Po drugie, dalsze zróżnicowanie jest wynikiem różnorodności ofert, produktów ubezpieczeniowych oraz docelowych grup klientów.

Ta różnorodność może być dodatkowo wzmocniona przez włączenie do zbioru przedmiotów badania witryn funkcjonujących na innych rynkach geograficznych. Aby zapewnić jak najlepszą reprezentatywność otrzymanych wyników, taka możliwość została wzięta pod uwagę. Po wstępnej analizie wybranych rynków zagranicznych okazało się, że poza kwestiami szczegółowymi (uwarunkowania prawne etc.) oraz oczywiście różnicami językowymi, witryny te nie wniosłyby jako dodatkowe przedmioty badania istotnych czynników różnicujących[194].

U podstaw badania znajduje się założenie, że metody stosowane przy opracowywaniu produktów ubezpieczeniowych są w swojej istocie ogólne i nie różnią się istotnie w poszczególnych krajach. Nie ma potrzeby robienia podobnego założenia dotyczącego funkcjonowania witryn internetowych od strony technologii, ponieważ jednorodność w tym zakresie gwarantowana jest przez znaczący i ciągle postępujący stopień standaryzacji w zakresie protokołów, formalizmów oraz technik wykorzystywanych w ich działaniu. Oczywiście istotne różnice ze względu na kryterium geograficzne stanowią prezentowane treści, kultura i lokalna specyfika rynku, szczegółowe potrzeby konsumentów, a także poziom kultury technicznej, a co za tym idzie, powszechność potencjalnych źródeł wiedzy ubezpieczeniowej.

Przy rozpatrywaniu omawianego podziału należy także zastanowić się nad sensownością definicji podziału geograficznego w sytuacji badania, którego obiektem są byty wirtualne, a zatem znajdujące się w przestrzeni informacji odrębnej w zasadniczy sposób od przestrzeni fizycznej. W związku z tym zastrzeżeniem wprowadzono następujące uszczegółowienie: przez witrynę przynależną do rynku danego kraju rozumiemy witryny internetowe prezentujące istotne treści w języku narodowym danego kraju z przeznaczeniem do sprzedaży dla klientów będących rezydentami danego kraju; w szczególności prowadzone przez krajowe podmioty. Ostatni warunek nie jest,

[194] Część ostatecznie przebadanych witryn ubezpieczeniowych w gruncie rzeczy należy do dużych grup ubezpieczeniowych prowadzących swoje biznesy w wielu krajach na terenie Europy. Ze względu na standardy korporacyjne oraz koszt infrastruktury IT grupy takie dążą do ujednolicania rozwiązań także w zakresie sprzedaży pomiędzy poszczególnymi rynkami, na których prowadzą operacje.

rzecz jasna, konieczny do spełnienia, aby zgodnie z definicją zaliczyć źródło wiedzy ubezpieczeniowej jako przynależne do konkretnego krajowego rynku.

Ze względu na uwarunkowanie, zaobserwowanie mało istotnych różnic w przeanalizowanych witrynach zagranicznych w stosunku do witryn rynku polskiego oraz fakt, iż badanie przeprowadzane jest w warunkach gospodarki polskiej jako części składowej europejskiego rynku ubezpieczeniowego określono, że przedmiotem badania są strony należące do polskich podmiotów.

6.2 Reprezentacja strukturalna źródła

Na całkowity opis źródła wiedzy składa się omówiona poniżej reprezentacja struktury oraz przedstawione w podrozdziale 6.3 rozszerzenia semantyczne. Opis taki ostatecznie przechowywany jest w postaci dokumentu XML, który zawiera informacje o jednym źródle. Formalna definicja zawartości takiego dokumentu została wyspecyfikowana w aneksie A („Język opisu procesu ekstrakcji").

Kompletna reprezentacja źródła składa się z następujących elementów:

1. deklaracji właściwości,
2. opisu wierzchołków odpowiadających elementom procesu nawigacji po źródle,
3. opisu skierowanego grafu łączącego zdefiniowane wierzchołki.

Poniżej przedstawiamy opis poszczególnych elementów.

6.2.1 Deklaracja właściwości

Wszystkie właściwości wykorzystane w procesie nawigacji po źródle są definiowane w ramach deklaracji tzw. abstrakcyjnych kontenerów właściwości. Praktycznie zawsze konkretyzacją takiego kontenera jest początkowy wierzchołek grafu nawigacji (Element startowy)[195]. Definicja właściwości składa się z trójki:

$$< s, W, c >, \tag{17}$$

gdzie:

- s jest jednoznacznym i unikalnym w całym grafie identyfikatorem właściwości,

[195] W niektórych przypadkach konieczne jest jednak deklarowanie dodatkowych kontenerów, zawierających niezależny podzbiór właściwości.

- W jest zbiorem dopuszczalnych wartości,
- natomiast $c \in W \cup \emptyset$ jest wartością początkową właściwości.

Dopuszczalny zbiór wartości definiowany jest w zależności od typu właściwości. W przypadku typów wyliczanych każda wartość w zbiorze deklarowana jest w sposób jawny. W przypadku typów interwałowych deklarowany jest przedział wartości za pomocą trójki:

$$< min, max, t >, \tag{18}$$

gdzie:

- *min* – wartość minimalna,
- *max* – wartość maksymalna,
- *t* – stała różnica pomiędzy kolejnymi wartościami zbioru wartości.

6.2.2 Właściwości warunkowe

W celu uelastycznienia możliwości definiowania zbiorów wartości przyjmowanych przez poszczególne właściwości oraz zwiększenia precyzji sterowania tymi wartościami, przewidziano szereg dodatkowych mechanizmów, wpływających zarówno na deklaracje zbiorów wartości *W*, jak też na przyjmowane przez właściwości wartości (w tym wartość początkową *c*)[196].

Właściwości warunkowe posiadają unikalną cechę możliwości zmiany zbioru przyjmowanych wartości w zależności od spełnienia dowolnie długiej listy warunków. Właściwości takie mają strukturę drzewa, którego wierzchołki mogą reprezentować albo elementy zbioru wartości właściwości albo warunki[197]. Zbiór możliwych wartości właściwości jest sumą elementów zbioru wartości zapisanych w tych wierzchołkach, dla których wierzchołki nadrzędne zawierają wyłącznie spełnione (prawdziwe) warunki. Warunki mają postać:

$$< s_w, pred, c_x >, \tag{19}$$

gdzie:

[196] Poza omówionymi właściwościami warunkowymi zaprojektowano także dwa inne mechanizmy, w tym: formalizm umożliwiający relatywizację przyjmowanych zbiorów wartości oraz wartości początkowych w zależności od czynników zewnętrznych, np. czasu; innym mechanizmem jest mechanizm łączenia lub wyodrębniania fragmentów wartości przyjmowanych przez inne właściwości – wartość właściwości w tym wypadku traktowana jest jak ciąg tekstowy.

[197] Warunki znajdujące się w tym samym elemencie traktowane są jako połączone operatorem logicznej sumy natomiast warunki znajdujące się w wierzchołkach podrzędnych połączone są operatorem logicznego iloczynu.

- s_w to jednoznaczny w całym grafie identyfikator właściwości,
- *pred* – jeden z predefiniowanych predykatów,
- c_x – stała reprezentująca wartość.

Predykat ma charakter prostych (atomowych) i zawsze dwuargumentowych relacji matematycznych. s_w i c_x stanowią argumenty dla predykatu.

6.2.3 Opis wierzchołków odpowiadających elementom procesu nawigacji

Zasadniczą część opisu procesu nawigacji i ekstrakcji zajmuje definicja wierzchołków grafu nawigacji. Kolejność definiowanych wierzchołków jest nieistotna, a wszelkie ewentualne referencje pomiędzy elementami odbywają się za pomocą jednoznacznych i unikalnych w ramach opisu źródła identyfikatorów, jakimi opatrzony jest każdy wierzchołek.

W podrozdziale 5.3 zestawiono w tabeli 7 możliwe elementy opisu procesu nawigacji i ekstrakcji. W zestawieniu tym wyróżniono przede wszystkim te elementy, które stanowią autonomiczne wierzchołki grafu procesu ekstrakcji. Obecnie zostaną one omówione.

Element startowy (Source) oraz Terminator są typami wierzchołków granicznych w grafie. Element startowy oznacza wierzchołek, który inicjuje proces nawigacji i ekstrakcji. Z kolei Terminator jest wierzchołkiem kończącym przejście przez graf. Przejście przez graf oznacza zakończenie pojedynczego, pełnego cyklu nawigacji. Proces ekstrakcji składa się z wielu cykli nawigacji po źródle. Stąd osiągnięcie wierzchołka terminalnego oznacza cofnięcie się automatu realizującego proces do wierzchołka startowego, a następnie sprawdzenie warunku stopu. Jeżeli warunek ten nie został osiągnięty, rozpoczyna się kolejny cykl nawigacji.

Element startowy, poza wyznaczeniem punktu zapoczątkowania cyklu procesu, jest jednocześnie kontenerem przechowującym definicje właściwości wykorzystywanych w procesie nawigacji i ekstrakcji oraz aktualnie przypisane im wartości. Z racji pełnionych ról zakłada się, że w poprawnym grafie istnieje tylko jeden punkt początkowy. Dozwolona jest dowolna liczba Terminatorów, aczkolwiek w typowych scenariuszach ekstrakcji także i ten typ wierzchołka ma pojedynczą reprezentację.

Ekstraktory to typ wierzchołków odpowiadających procesowi ekstrakcji danych z aktualnie przetwarzanej, tj. w miejscu wystąpienia wierzchołka w grafie, treści pochodzącej ze źródła. Zadaniem abstrakcyjnego ekstraktora jest wykonanie konkretnej procedury ekstrakcji na podstawie informacji dostarczonej w definicji wierzchołka. Procedura ekstrakcji polega na zlokalizowaniu użytecznych danych oraz przepisaniu ich jako wartości do jednej lub wielu właściwości. Tak przypisane wartości są dowolnie osiągalne i wykorzystywane przez automat wykonujący proces nawigacji w trakcie przetwarzania kolejnych wierzchołków grafu.

Zgodnie ze stanem zaprezentowanym w tabeli 7 przedstawiającej typy wierzchołków, w prototypie zaimplementowano trzy rodzaje ekstraktorów wykorzystujących różne mechanizmy i reguły ekstrakcji. Są to:

- ekstraktor wyrażeń regularnych – wykorzystuje mechanizm wyrażeń regularnych do przeszukiwania strumienia danych pochodzących ze źródła w celu wykrycia występowania dowolnej liczby wzorców; identyfikatory właściwości zaszyte są w oznaczenia grup przechwytywania (capture group),
- ekstraktor języka XPath – zawiera listę zapytań w formalizmie XPath, które są sekwencyjnie wykonywane na strumieniu (zazwyczaj w języku (X)HTML) z serwera; każde zapytanie jest skojarzone z jednym lub większą liczbą identyfikatorów właściwości,
- ekstraktor Przekształcenia Rozszerzalnego Języka Arkuszy Stylów (XSLT)[198] – zawiera skrypt przekształceń, który jest uruchamiany z wykorzystaniem procesora XSLT na dokumencie pochodzącym ze źródła; skrypt ma za zadanie wygenerować ciągi par postaci <identyfikator właściwości, wartość>.

Wzorce nawigacji jako podstawowa struktura nawigacji w prezentowanym prototypie zostały przedstawione w podrozdziale 5.4. Element tego typu reprezentuje pojedyncze przejście pomiędzy podstronami w procesie nawigacji. Stąd w grafie znajdują się one wszędzie tam, gdzie w procesie nawigacji następuje wczytanie nowego doku-

[198] Podobny mechanizm ekstrakcji informacji zastosowano w [Węcel2011].

mentu z adresu URL za pomocą metody GET oraz w części przypadków wykorzystania metody POST[199].

Automatyzujące wzorce nawigacji stanowią rozszerzenie wzorców nawigacji. Istotna różnica wpływająca na decyzję o utworzeniu odrębnego typu polega na dodatkowym przechowywaniu przez ten typ wzorca nawigacji informacji o symulowanych działaniach potencjalnego użytkownika wykonywanych na aktualnie przetwarzanym dokumencie. W praktyce oznacza to, że automatyzujące wzorce nawigacji opisują poszczególne formularze generowane przez źródło webowe. Opis formularza składa się z drzewa czynności (AutomationTask), jakie wykonywałby potencjalny użytkownik, gdyby jego celem było poprawne wypełnienie istotnych informacji na formularzu oraz jego odesłanie na serwer.

Symulacja wykonania poszczególnych czynności w drzewie ma charakter warunkowy. Oznacza to, że formalizm opisu tego typu elementów przewiduje możliwość wykonania określonego podciągu czynności wtedy i tylko wtedy, gdy prawdziwy jest określony warunek. Sposób deklarowania warunków jest analogiczny do zapisu stosowanego w przypadku omówionych w podrozdziale 5.4 warunkowych zbiorów wartości właściwości.

W celu poprawy jakości symulacji użytkownika oraz eliminacji ewentualnych problemów w komunikacji pomiędzy klientem webowym a serwerem, każda z czynności wykonywanych w sekwencji oddzielona jest w czasie za pomocą jednej z kilku możliwych do wyboru implementacji miernika czasu (AbstractTimer). Wybór implementacji oraz szczegóły konfiguracji mierników czasu używanych w danym drzewie czynności zależą od definicji wierzchołka. W szczególnym przypadku, jeżeli z definicji wierzchołka wynika, że formularz, który jest przez niego opisywany ma charakter dynamiczny, tj. zawiera pola, których wartość jest w trakcie wypełniania modyfikowana lub zależna od innych pól, wówczas możliwy jest do wykorzystania specjalny miernik czasu nadzorujący proces wymiany dodatkowych danych z serwerem, który zwraca sterowanie dopiero po zakończeniu obsługi asynchronicznego żądania HTTP(S).

[199] O wykorzystaniu obydwu metod w kontekście komunikacji serwera z przeglądarką internetową wspominamy w podrozdziale 2.1.2.

W tabeli 9 zestawiono rodzaje czynności obsługiwane przez automatyzujące wzorce nawigacji w charakterze symulacji czynności użytkownika.

Tabela 9. Rodzaje czynności obsługiwane przez automatyzujące wzorce nawigacji
Źródło: opracowanie własne

Czynność	Argumenty	Uwagi
Nawigacja za pomocą hiperłącza	URL lub nazwa lub identyfikator węzła HTML	W przypadku nazwy lub identyfikatora poprzez metodę click()
Naciśnięcie przycisku (ClickButton)	Nazwa lub identyfikator węzła HTML	W przypadku przycisku typu „submit" może służyć do zatwierdzenia formularza
Zaznaczenie lub zwolnienie dwustanowego przycisku zaznaczenia (ToggleCheckBox)	Nazwa lub identyfikator węzła HTML, stan	
Wprowadzenie danych do pola tekstowego (EnterData)	Nazwa lub identyfikator węzła HTML, wartość wpisywana	
Wypełnienie obszaru tekstowego (EnterDataTextArea)	Nazwa lub identyfikator węzła HTML, wartość wpisywana	
Wybór elementu z listy (SelectListItem)	Nazwa lub identyfikator węzła HTML listy, zbiór zaznaczonych wartości lub ich prezentowanych odpowiedników	
Akceptacja aktualnego formularza (SubmitForm)	Nazwa lub identyfikator węzła HTML	Poprzez wywołanie metody submit()
Zaznaczenie lub zwolnienie jednego z serii przycisków wyboru alternatywy (SelectRadioButton)	Nazwa lub identyfikator węzła HTML, stan (opcjonalnie)	
Zaznaczenie lub zwolnienie jednego z serii przycisków wyboru alternatywy (SelectRadioButton2)	Nazwa lub identyfikator węzła HTML, wartość odpowiadająca przyciskowi	

Elementy zrzucające (Dumper) reprezentują w grafie węzły wynikowe. Ich zadaniem jest przygotowanie i sformatowanie zbieranych przez automat nawigujący informacji oraz wydelegowanie pojedynczo lub w zagregowanej formie poza system ekstrakcji. W tym celu element zrzucający może wykorzystać urządzenie wejścia-

wyjścia lub dowolne dostępne repozytorium danych. W prezentowanym prototypie zaimplementowano dwa rodzaje elementów zrzucających:

- generujący plik w formacie CSV[200],
- zapisujące dane w wybranej bazie danych[201].

Element zrzucający jest zasadniczo umiejscawiany na końcu cyklu nawigacji, jeżeli przetwarzany jest w każdym cyklu. W rzadkich przypadkach może być przetworzony raz w całym procesie ekstrakcji. Teoretycznie możliwa jest konfiguracja grafu, w której występuje więcej niż jeden element zrzucający. Niemniej, takie rozwiązanie wydaje się mieć ograniczone zastosowanie praktyczne.

6.2.4 Opis grafu nawigacji

Trzecim elementem reprezentacji struktury źródła jest opis grafu nawigacji. Graf nawigacji przedstawić można jako uporządkowana para:

$$G := < V, A >, \tag{20}$$

gdzie:

- V to zbiór wierzchołków opatrzonych identyfikatorami,
- A to zbiór uporządkowanych par będących krawędziami skierowanymi.

Ten element opisu źródła sprowadza się do listy deklaracji powiązań pomiędzy dwoma wierzchołkami wskazanymi za pomocą identyfikatorów oraz nadania identyfikatora dla tak utworzonego wierzchołka. Minimalny poprawny graf składa się z dwóch elementów krańcowych (startowego oraz terminalnego) oraz jednego wzorca nawigacji połączonych dwoma krawędziami.

Składnia języka opisu grafu nie nakłada żadnych ograniczeń dla tworzonych połączeń pomiędzy poszczególnymi typami wierzchołków. Niemniej, na poziomie semantycznym, nie każda dowolna krawędź jest sensowna, a w związku z tym prawidłowa[202]. W szczególności dotyczy to także kierunku krawędzi.

[200] ang. Comma Separated Values. Standardowy plik z danymi tabelarycznymi. Pliki tego typu można odczytać i edytować za pomocą popularnego oprogramowania biurowego.
[201] W celu obsługi niskopoziomowego bazy danych wybrano ramę ORM nHibernate. Takie rozwiązanie zasadniczo zwiększa elastyczność oraz pozwala użyć praktycznie dowolny silnik zarządzania bazą relacyjną bez konieczności jego oprogramowania.
[202] Prototypowe rozwiązanie w tym zakresie nie ma żadnego szczególnego mechanizmu kontroli. Ewentualne błędy wykrywane są dopiero w trakcie realizacji procesu nawigacji.

Przedstawione powyżej (oraz w podrozdziale 6.3) elementy reprezentacji źródła internetowego mają bezpośredni wpływ na sposób działania mechanizmów nawigacji oraz ekstrakcji danych. Stanowią one adaptację, modyfikację lub ulepszenie analogicznych, najbardziej zaawansowanych rozwiązań opisanych w podrozdziale 2.2. W szczególności do istotnych ulepszeń zaliczyć należy:

- semantyczne rozszerzenie opisu,
- sterowanie procesem wspomagane ontologią,
- koncentrację na trudnościach w obsłudze dynamicznych treści witryn internetowych,
- rozwinięcie struktur oraz elementów kontroli poprawiających możliwość obsługi dynamicznych treści witryn internetowych,
- wprowadzenie właściwości, stanowiących jednocześnie element wynikowy procesu oraz (pośrednio) element sterujący.

6.3 Reprezentacja semantyczna – model struktury wiedzy

Zgodnie z tabelą 6 podaną w podrozdziale 5.3 jedną z istotnych decyzji projektowych przy opracowywaniu metody oraz bazującego na niej prototypu było zastosowanie zewnętrznej struktury wiedzy dla poprawy kontroli sterowania procesem nawigacji i ekstrakcji ze źródeł internetowych.

Wiedza zapisana w ontologii pozwala na zdefiniowanie standardowych symboli[203], przypisanie im określonych znaczeń oraz zdefiniowanie relacji zachodzących pomiędzy tymi symbolami. Istnieje szereg przesłanek przemawiających za wykorzystaniem standardowych formalizmów opisu wiedzy w prezentowanej metodzie. Przede wszystkim korzyść przynosi możliwość wykorzystania gotowych standardów oraz wsparcie przeznaczonych dla nich narzędzi, a w przypadku standardów opartych na XML dochodzi jeszcze łatwość integracji z dokumentami zawierającymi informację w innych formatach webowych. Pozwala to również podnieść ogólność rozwiązania poprzez modularność i łatwa rozszerzalność. Kolejną przesłanką jest potrzeba przechowania informacji o zaistnieniu dodatkowych faktów związanych z reprezentacją źródeł. Należy również wspomnieć utrudnione modelowanie wiedzy przy zastoso-

[203] Przez „symbol” należy rozumieć tutaj jednoznaczny i unikalny ciąg znaków, tworzący nazwę (pojęcie).

waniu współczesnych języków programowania (przykładowo brak wsparcia dla wielodziedziczenia[204]), co nie stanowi problemy w przypadku podejścia ontologicznego.

Mechanizm reprezentacji semantycznej oraz wsparcia przy jej zastosowaniu procesu ekstrakcji jest następujący. W pierwszej kolejności (faza wsparcia tworzenia opisu źródła), w przypadkach kiedy jest to niezbędne, pojęciami z ontologii oznaczane są elementy poszczególnych podstron lub całych stron. Następnie pojęcia z poziomu warstwy dokumentów (X)HTML są przenoszone na poziom reprezentacji struktury źródła. W niektórych przypadkach konieczna jest dalsza, manualna adnotacja opisu źródła. W dalszej kolejności – już w fazie nawigowania po źródle przez automat - dokonuje on analizy wystąpienia dodatkowych oznaczeń pojęciami związanych z przetwarzanym elementem. W takiej sytuacji automat wykonuje wnioskowanie odnośnie do relacji pomiędzy pojęciami. Jeżeli zostały ustalone odpowiednie reguły, to automat zrealizuje niestandardowe zachowanie (w stosunku do normalnego trybu pracy).

Struktura ontologii podzielona została na 3 względnie niezależne i wydzielone subontologie[205] [Abramowicz2011]: subontologia produktu, subontologia ryzyk oraz subontologia czynników ryzyka.

Dodatkowe diagramy ilustrujące treści poruszane w podrozdziałach 6.3.1-6.3.3 zamieszczono w aneksie B („Ontologia”).

6.3.1 Subontologia produktu

Pojęcie produktu ubezpieczeniowego zostało szeroko omówione w podrozdziale 3.1. Istotą tej części modelu wiedzy jest opis aspektów prawnych oraz marketingowych związanych z ofertą ubezpieczenia. Ze względu na wyznaczony zakres badań oraz prezentowaną metodę, aspekty prawne nie należą do przedmiotu zainteresowania niniejszej pracy[206].

Pierwotnie w przyjętym rozwiązaniu założono[207], że produkt ubezpieczeniowy (nie występujący jako niezależny byt) modelowany jest za pomocą współwystępują-

[204] Problem ten można co prawda próbować rozwiązać na różne sposoby, tj. za pomocą pewnych sztuczek programistycznych oraz określonych wzorców projektowych. Jednakże jest wątpliwe, aby korzyści z takiego podejścia przewyższyły te związane z użyciem gotowych rozwiązań, na rzecz których optujemy.
[205] Przedrostek „sub” oznacza tutaj element podrzędny, strukturę stanowiącą część struktury wyższego rzędu.
[206] Mimo to aspekty te posiadają istotny potencjał w zakresie wykorzystania w przypadku np. rozszerzenia prezentowanej metody o możliwość porównywania ofert z punktu widzenia zakresu ryzyk i odpowiedzialności.
[207] Pewne koncepcje przy konstrukcji modelu zaczerpnięto z [Kowalewski2006].

cych pojęć: Polisy, Ubezpieczenia[208] oraz UmowyUbezpieczeniowej. Z ograniczeń kardynalności zadeklarowanych w modelu wnioskować można, że istnienie instancji pierwszego z trzech pojęć nie jest warunkiem koniecznym dla występowania instancji Ubezpieczenia (rysunek 22).

Na całkowity zbiór instancji WarunkówUmowyUbezpieczeniowej składają się obiekty wchodzące w skład instancji klasy UmowaUbezpieczeniowa oraz OgólneWarunkiUbezpieczenia. Wśród instancji WarunkówUmowyUbezpieczeniowej osobną podklasę stanowią ElementyUmowyUbezpieczeniowej. To ostatnie pojęcie określa charakterystyczne warunki, które zgodnie z konstrukcją umowy ubezpieczenia powinny w niej obowiązkowo wystąpić. Przykładowo Składka jest pojęciem, którego instancja należy do zakresu (podklasa) pojęcia ElementyUmowyUbezpieczeniowej.

Regulacje umowne w przypadku modelowania ubezpieczenia muszą także wskazywać zakres odpowiedzialności, tj. zbiór obiektów pojęcia Ryzyko, od którego dane ubezpieczenie ma zapewniać ochronę. Wiedza o powiązaniu Ryzyk z WarunkamiUmowyUbezpieczenia odbywa się przez odpowiednie wykorzystanie konfiguracji dwóch pojęć powołanych do reprezentacji reguł odpowiedzialności umownej: ZasadyOdpowiedzialności, ZasadyWykluczeniaOdpowiedzialności, a także odpowiednio powiązanymi z nimi pojęć WłączeniaOdpowiedzialności lub WyłączeniaOdpowiedzialności. Dokonując analizy tego fragmentu modelu dziedziny, wydać się może nieintuicyjne rozwiązanie polegające na tym, że zdefiniowano ZasadyWykluczeniaOdpowiedzialności na bazie pojęcia ZasadyOdpowiedzialności, jako że wydają się one przeciwstawne. Za takim rozwiązaniem przemawia jednak fakt, że współdzielą one prawie wszystkie właściwości. W prezentowanej metodzie ta subontologia będzie przykładowo pomocna w odpowiedniej interpretacji sytuacji związanych z różnorodnymi opcjami płatności (np. zniżki, raty etc.)

6.3.2 Subontologia ryzyk

Omawiając poprzednią część modelu wiedzy wykorzystanego w procesie badawczym, odwołano się do pojęcia Ryzyka. Mimo tego w subontologii produktu nie zaj-

[208] Celowo unikamy nazwania tego konceptu RodzajemUbezpieczenia ze względu na przyjętą konwencję modelowania. RodzajemUbezpieczenia nazwalibyśmy meta-klasę, której instancjami byłyby pojęcia reprezentujące poszczególne rodzaje ubezpieczeń.

mowaliśmy się, ani nie zakładaliśmy żadnej konkretnej definicji tego pojęcia. Podejście takie było celowe ze względu na jak największe odseparowanie poszczególnych modeli składowych współtworzących całą ontologię.

Dokonując przeglądu literatury zauważyć trzeba, że jest to jedno z centralnych pojęć we współczesnej teorii ubezpieczeń. Badacze zajmujący się ubezpieczeniami wyróżniają szereg definicji ryzyka. Kolejne definicje uwypuklają poszczególne przejawy tego pojęcia[209], a przez to także specyficzne konteksty użycia. Mimo iż w opisywanym badaniu pojęcie to nie pełni roli pierwszoplanowej, to przeprowadzamy skrótową dyskusję dotyczącą tego zagadnienia ze względu na wagę konceptu dla modelu domeny jako takiego. Można bowiem zaryzykować stwierdzenie, iż bez zaistnienia zespołu zjawisk składających się na to pojęcie – bez względu na wybraną precyzyjną definicję lub model – cała dziedzina ubezpieczeń nie miałaby racji bytu.

Poniżej dokonujemy przeglądu zaproponowanych schematów modelowania ryzyka[210] oraz podajemy przesłanki, którymi kierowano się podejmując decyzję co do wybranego rozwiązania.

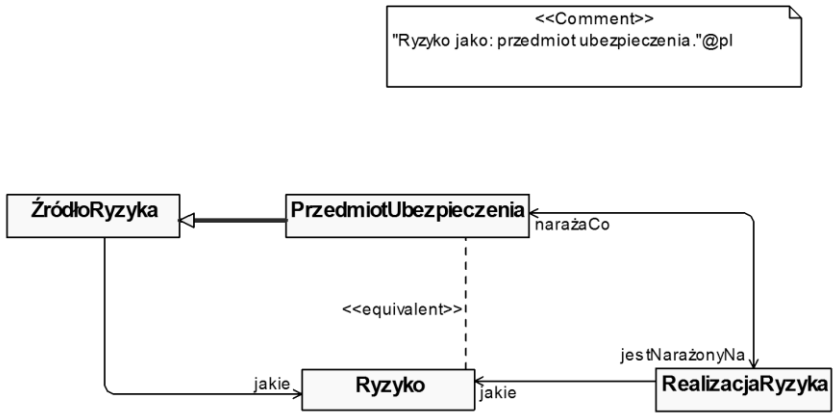
Wszystkie z przytoczonych schematów modelujących zasadniczo różnią się między sobą. Mimo to można dla nich doszukać się także pewnych cech wspólnych. Do takich cech zaliczyć można koncentrację raczej na skali mikro zagadnień związanych z wystąpieniem ryzyka niż podejście od strony organizacji ubezpieczeniowej jako wyodrębnionej całości. Inną cechą jest nieuwzględnianie zmienności w zjawisku jakim jest ryzyko. Szablony te, w przedstawionej formie, nie uwzględniają również istnienia związków pomiędzy danymi statystycznymi oraz wiedzą o przedmiocie ubezpieczenia. Z punktu widzenia przydatności w przedstawionym badaniu takie uproszczenia nie są dyskwalifikujące przy rozważeniu wyboru szablonu, aczkolwiek stanowią pewien rodzaj ułomności.

Pierwszy z przytoczonych wzorców stanowi próbę zamodelowania pragmatycznej definicji Ryzyka, często przytaczanej w praktyce biznesu ubezpieczeniowego[211].

[209] Fakt istnienia różnych definicji wynika z trudności w ujęciu licznych aspektów tego pojęcia w jednorodny i całościowy sposób.
[210] Przedstawiamy 5 modeli odpowiadających różnym definicjom pochodzącym z opracowań naukowych dotyczących ryzyka w ubezpieczeniach [Monkiewicz2000], [Michalski2004].
[211] Modelowana definicja jest krytykowana ze strony znawców teorii ubezpieczeń. Mimo to była brana pod uwagę przy tworzeniu ontologii na potrzeby badawcze.

W definicji tej za równoważne uznaje się pojęcia Ryzyka i PrzedmiotuUbezpieczenia (rysunek 10). Dalsze semantyczne implikacje będące pochodną wymienionej równoważności oznaczają, że PrzedmiotUbezpieczenia można także uznać za ŹródłoRyzyka. Ponadto jest on także narażony na RealizacjęRyzyka.

Zarówno model, jak i definicja, na bazie której został on zbudowany, powoduje pojawienie się istotnych zastrzeżeń z punktu widzenia stosowania dobrych praktyk inżynierii wiedzy. Wynika to zasadniczo z faktu, iż definicja pojęcia Ryzyka jest bardzo daleko posuniętym skrótem myślowym, wygodnym do stosowania podczas komunikacji przy realizacji części procesów dystrybucji. Brak precyzji cechujący tę komunikację na poziomie formalnej specyfikacji objawia się jednak potencjalnymi błędami logicznymi[212]. W rezultacie szablon ten jest nie do zaakceptowania.

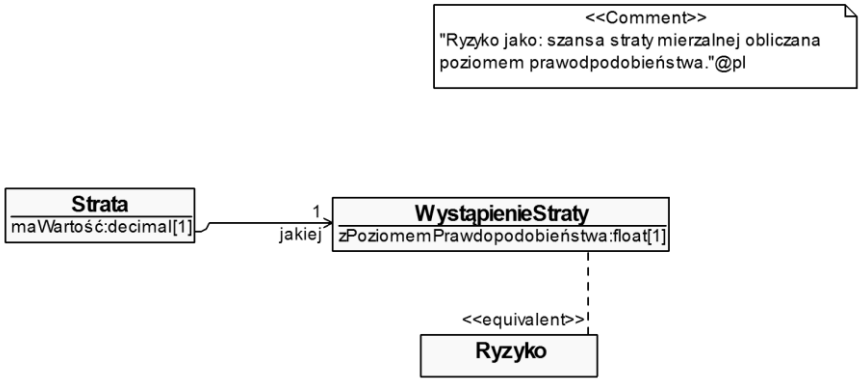


Rysunek 10. Model UML pojęcia "Ryzyko" w postaci definicji pragmatycznej
Źródło: opracowanie własne

W drugim z omówionych podejść (rysunek 11) Ryzyko jest pojęciem równoważnym z terminem WystąpienieStraty. Strata ma cechę mierzalności, co oznacza, że może być wyrażona w określonym mierniku wartości. WystąpienieStraty jest jednak tożsame z pojęciem Ryzyka tylko wówczas, jeżeli zdarzenie je powodujące może potencjalnie wystąpić i jeszcze nie zaszło. Takie obostrzenia nie są wprost reprezentowane w rozważanym wzorcu. Częściowo rekompensuje tę niedoskonałość wprowadzenie pomocniczej umownej miary prawdopodobieństwa. Zaletą takiego schematu jest oszczędność w zakresie wykorzystania modelujących bytów, a co za tym idzie, prostota. Jednocześnie jest to także istotną wadą, ponieważ szablon ten bardzo spłyca interpre-

[212] Można dojść do wniosku, że Ryzyko jest ŹródłemRyzyka.

tację analizowanego pojęcia, przez co nie pozwala brać pod uwagę szeregu ważnych czynników, które w modelu wiedzy dziedzinowej powinny zostać uwzględnione. Postulat większej szczegółowości uwzględniają natomiast pozostałe szablony.



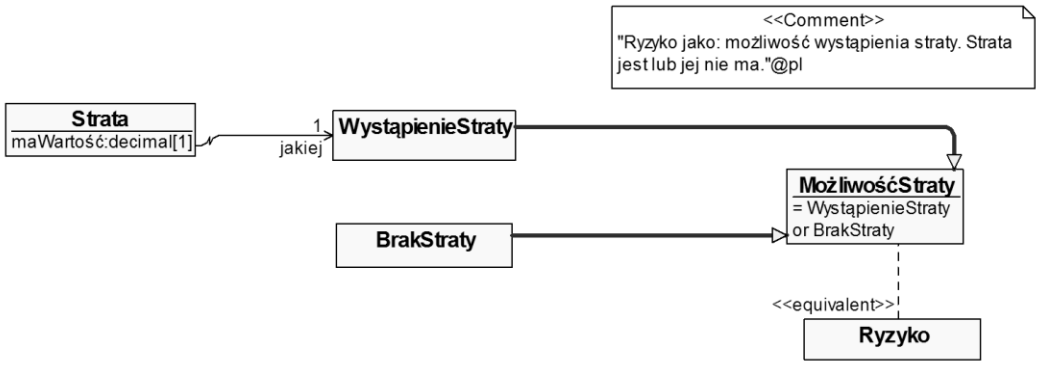
Rysunek 11. Model UML pojęcia "Ryzyko" definiowanego poprzez mierzalną stratę
Źródło: opracowanie własne

Następne z analizowanych rozwiązań wykazuje podobieństwo do poprzedniego modelu, w tym że wykorzystuje ono także termin WystąpienieStraty (rysunek 12). Nie jest to jednak tym razem koncept centralny dla całego schematu. Pojęciem ekwiwalentnym dla Ryzyka jest nowo wprowadzony termin MożliwośćStraty. Ten ostatni ma za zadanie odzwierciedlenie dwustanowości wystąpienia straty wprowadzonej przez definicję, na podstawie której szablon modelujący powstał. MożliwośćStraty może w ten sposób przyjąć jeden z dwóch stanów: strata wystąpiła albo nie. Drugie podobieństwo, jakie wykazuje obecnie dyskutowany wzorzec w stosunku do poprzednika, to kopia pojęcia Straty. Termin ten może być rozumiany w związku z tym analogicznie. W stosunku do pierwszego modelu rozwiązano tutaj ułomność pojęcia WystąpienieStraty – szablon *explicite* reprezentuje możliwe stany rzeczywistości. Elementem uwstecznającym z kolei jest fakt, że schemat nie zawiera odniesienia do probabilistycznej natury zjawisk[213].

Kolejny kandydujący do wyboru wzorzec zasadniczo niweluje ujemne aspekty dwóch poprzednich. Przede wszystkim, w stosunku do poprzedniego, koncentruje się na rozkładach prawdopodobieństwa (rysunek 13). Aby skutecznie to zrobić, w szablonie wprowadzono pomocnicze pojęcia abstrakcyjnego Wyniku oraz stanowiących jego

[213] Aczkolwiek obie z modelowanych definicje wykazują podobieństwo, to w literaturze przedmiotu właśnie ta okoliczność podnoszona jest jako wyróżniająca obydwa sposoby spojrzenia na rozumienie ryzyka.

konkretyzację WynikuOczekiwanego[214] oraz WynikuInnego. Każdy Wynik jest w tym modelu Zdarzeniem, co stanowi dopuszczalne, semantyczne uproszczenie związku przyczynowo-skutkowego. Wreszcie w schemacie tym dla każdej instancji Wyniku przypisana jest odpowiednia informacja o prawdopodobieństwie za pomocą obiektu będącego instancją PrawdopodobieństwaWyniku. Ryzyko jest odpowiednikiem szczególnego rodzaju tego prawdopodobieństwa, mianowicie przypisanego do każdej instancji WynikuInnego. W rezultacie przyjęcia takiej interpretacji Ryzyko jest tożsame z konceptem PrawdopodobieństwoWynikuInnego (od oczekiwanego).

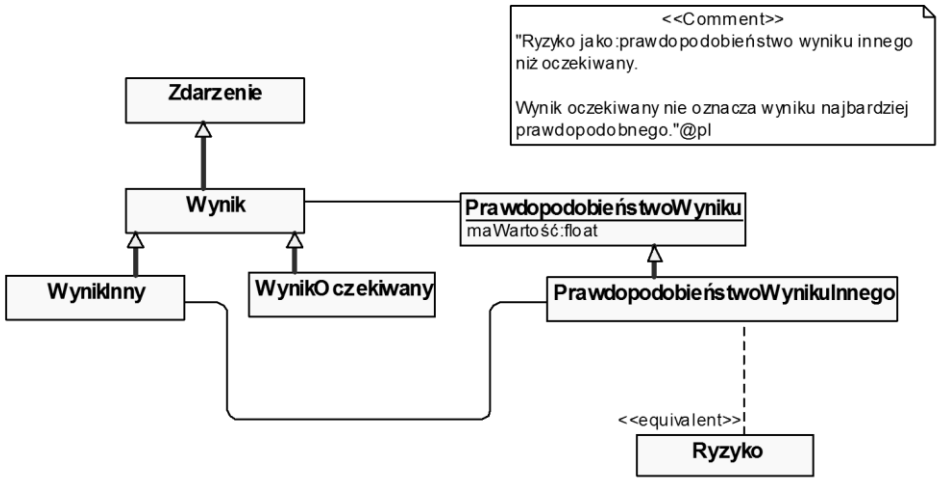


Rysunek 12. Model UML pojęcia "Ryzyko" rozumianego jako możliwość straty
Źródło: opracowanie własne

Analizowany schemat w stosunku do wszystkich pozostałych kandydatów jest najbardziej rozbudowany. Ponadto dobrze obrazuje naturę powiązań pomiędzy takimi kluczowymi dla domeny konceptami jak Zdarzenie, Wynik oraz natury niepewności. Wzorzec generuje jednak dwa potencjalne problemy w przypadku praktycznego zastosowania. Po pierwsze, w takim schemacie każde możliwe do zaistnienia zdarzenie jest źródłem wyniku, które w większości stanowić będą instancję WynikuInnego[215]. Każdemu takiemu wynikowi przypisać można pewne prawdopodobieństwo – trudno jednak wskazać na kryteria jego wyznaczenia. Dodatkowo oznacza to istnienie bardzo wielu bytów typu PrawdopodobieństwoWynikuInnego, a zatem z każdym z nich trzeba by łączyć osobne Ryzyko. Po drugie, Zdarzenia jako uogólnienie wcześniej pojawiającego się pojęcia Straty mogą mieć także pozytywne implikacje. Oznacza to,

[214] Wynik oczekiwany może być tutaj rozumiany zazwyczaj jako neutralny wariant rzeczywistości, czyli taki, w którym nie dochodzi do niezakładanego (a zatem najczęściej negatywnego) zdarzenia.
[215] Najczęściej odpowiadającego sytuacji straty lub wypadku etc.

że taki wzorzec wymusza jednocześnie uwzględnianie Ryzyka powiązanego z pożądanymi stanami rzeczywistości[216].



Rysunek 13. Model UML pojęcia "Ryzyko" – prawdopodobieństwo nieoczekiwanego wyniku
Źródło: opracowanie własne

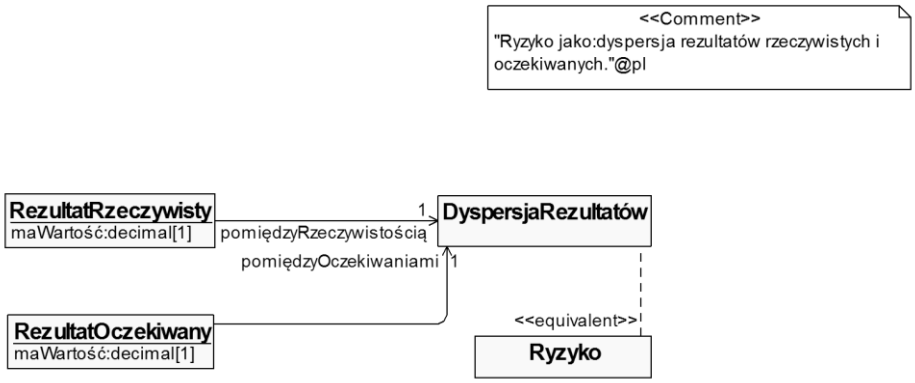
Ostatni z kandydatów rozważanych przy wyborze wzorca reprezentującego pojęcie Ryzyka zaprezentowano na rysunku 14. W tym ujęciu pojęcie Ryzyka odpowiada pojęciu DyspersjiRezultatów. Dyspersja ta stanowi różnicę pomiędzy RezultatemRzeczywistym a RezultatemOczekiwanym. Takie rozwiązanie oznacza, że dla każdego obiektu uznawanego za RezultatRzeczywisty istnieć musi analog w postaci instancji pojęcia RezultatOczekiwany. Na potrzeby projektowanego wzorca założono, że oba rodzaje rezultatów są mierzone w wartościach pieniężnych.

Zaletą tego wzorca jest niski stopień skomplikowania. W odróżnieniu od innych schematów zakłada on pomiar ryzyka w jednostce pieniężnej. Jeśli chodzi o wady, w stosunku do poprzedniego wzorca, to aktualnie analizowany szablon nie reprezentuje powiązań w sieci znaczeniowej ważnych, jak się wydaje, dla teorii ubezpieczeń pojęć.

Wybór odpowiedniego kandydata spośród przedstawionych szablonów oraz definicji wpływa w sposób wyraźny na rozwiązania implementacyjne i funkcjonowanie całej ontologii. Dzieje się tak dlatego, że poszczególne subontologie wymagają integracji przez połączenie pojęć w nich zawartych różnymi relacjami z innymi elemen-

[216] Konsekwencją takiego stanu rzeczy jest możliwość pojawienia się dychotomii na ryzyko klienta oraz ryzyko ubezpieczyciela – co niekoniecznie jest pożądane.

tami pozostałych części. Zaadoptowany wzorzec musi przede wszystkim odzwierciedlać cele postawione przed konstruowanym modelem jako całością.



Rysunek 14. Model UML pojęcia "Ryzyko" rozumiany jako dyspersja rezultatów
Źródło: opracowanie własne

W przypadku tworzenia ontologii ogólnego przeznaczenia lub ontologii integrującej informacje różnego pochodzenia dopuszczających różnice w interpretacji pojęć, użytecznym rozwiązaniem będzie zastosowanie więcej niż jednego modelu jednocześnie.

Zaznaczyć także należy, że przedstawiono tutaj ilustrację dla najbardziej reprezentatywnych konstrukcji pojęciowych ryzyka. Lista definicji jest oczywiście bogatsza. Dla dalszych zastosowań subontologia ryzyk stanowić będzie dopełnienie umożliwiające modelowanie takich sytuacji, jak np. produkty zagregowane (chroniące przed różnymi ryzykami) oraz techniki cross-selling (produkty chroniące przed komplementarnymi ryzykami).

6.3.3 Subontologia czynników ryzyka

Podstawowym zadaniem dla tej wyodrębnionej części modelu wiedzy jest wsparcie w reprezentacji i opisie zmiennych taryfikacyjnych, zmiennych w algorytmie dopuszczenia do ubezpieczenia, a także odniesienie tych elementów modeli aktuarialnych do przechowywanej w prototypowym rozwiązaniu wiedzy o ryzykach.

Przy tworzeniu omawianej subontologii wykorzystano doświadczenia pochodzące z wcześniejszych prac, a w szczególności ze szkicu tego modelu wiedzy stworzonego na etapie wcześniejszych badań [Abramowicz2011]. Zastosowano wówczas pojęcie czynnika ryzyka, które zdefiniowano jako „czynnik wypływający na ryzyko; rozumia-

ny w formie dowolnego elementu w postaci informacji, który może mieć wpływ na wycenę ryzyka, a w konsekwencji na cenę produktu ubezpieczeniowego".

Pierwsza wersja omawianej obecnie subontologii przygotowana została na podstawie analizy 11 witryn internetowych prowadzących sprzedaż ubezpieczeń. W celu zbudowania modelu dokonano wtedy przeglądu procesu sprzedaży 17 odrębnych produktów. Mimo, że faza inżynierii wiedzy poparta była staraniami zmierzającymi do doboru reprezentatywnych przykładów produktów oraz zróżnicowanych źródeł internetowych oferujących ubezpieczenia on-line, trudno jest uznać stworzony model za wyczerpujący i w pełni zadowalający. Tym bardziej, że nie był on budowany dla realizacji żadnych szczególnych zadań, tak jak ma to miejsce w przypadku prezentowanego w niniejszej pracy badania.

Pierwotny model, którego rozszerzenie zostało dokonane dla zastosowania w prezentowanej pracy metodzie, przyjmował jako centralne pojęcie CzynnikRyzyka (rysunek 21). Jego opis został uzupełniony o informację, iż jest to faktyczny składnik uwzględniany w procesie wyznaczania składki. Takie uszczegółowienie każe traktować je zatem jako odpowiednik zmiennej taryfikacyjnej. W modelu wyróżniono jednak jednocześnie powiązane z instancjami CzynnikówRyzyka obiekty będące realizacją pojęcia ManifestacjaCzynnikaRyzyka. To ostatnie pojęcie należy traktować jako odzwierciedlenie danego czynnika – zmiennej w formularzu dostępnego dla ubezpieczanego lub ubezpieczającego. Przykładowo do kalkulacji składki potrzebny jest parametr określający liczbę mieszkańców w mieście, w którym znajduje się przedmiot ubezpieczenia, natomiast faktycznie wartość ta ustalana jest na podstawie kodu pocztowego. Instancją CzynnikaRyzyka jest liczba mieszkańców, podczas gdy Manifestacją tego czynnika jest kod pocztowy. W części przypadków manifestacje są identyczne z odpowiadającymi sobie czynnikami, co należy rozumieć jako sytuację, w której do instancji obydwu terminów przypisany jest dokładnie ten sam obiekt klasy ZbiórWartości. Taki układ oznacza, że dany CzynnikRyzyka wraz z przynależnym mu zbiorem wartości jest explicite przedstawiany do wskazania przez potencjalnego klienta, a następnie bez żadnych modyfikacji wykorzystywany w algorytmie taryfikacyjnym. Oczywiście model dopuszcza także sytuację alternatywną – zbiory wartości są różne lub w szczególnym przypadku zupełnie odmienne. W rezultacie CzynnikRyzyka

ma charakter ukryty, a jego wartość wyznaczana w procesie wyznaczenia taryfy oraz składki jest funkcją wartości przyjętej przez odpowiadającą ManifestacjaCzynnikaRyzyka.

W dalszej części referowany model wprowadza szereg podklas dla pojęcia CzynnikaRyzyka. Podklasy te wyróżniono na podstawie analizy formularzy stanowiących elementy modeli wyceny ubezpieczeń. W ten sposób wprowadzano podział na CzynnikiRyzykaWymagane oraz CzynnikiRyzykaNiewymagane. Pierwsze mają charakter stały i obowiązkowy w modelu wyceny składki źródła webowego, natomiast w drugim przypadku wprowadzenie informacji dotyczącej czynnika nie jest obligatoryjne – model może poprawnie funkcjonować mimo tego. Dodatkowo szczególnym rodzajem pojęcia CzynnikiRyzykaNiewymagane jest CzynnikiRyzykaNieistotne, czyli taki czynnik, który posiada swoją manifestację, ale nie jest brany pod uwagę w aktualnym modelu wyceny składki. Przykładem może być tu pole w formularzu dotyczące informacji o poprzednim ubezpieczycielu, które powinno być zidentyfikowane nie jako czynnik ryzyka tylko zmienna proceduralna, mimo że może mieć wpływ na kalkulację składki.

Inny podział koncentruje się na zbiorach przyjmowanych wartości. Z tego punktu widzenia w modelu wyróżniono dwa podzbiory. W zamierzeniu tworzyć je mają instancje pojęć: CzynnikRyzykaJakościowy oraz CzynnikRyzykaIlościowy. Te pierwsze mają charakter binarny lub wyliczeniowy, z kolei te drugie mają zdefiniowany przedział wartości liczbowych.

Jeszcze jednym przewidzianym przez zaproponowaną subontologię szczególnym rodzajem CzynnikaRyzyka jest CzynnikRyzykaWarunkowy. Jest to taki CzynnikRyzyka, który zmienia swój stan. W zależności od realizacji określonego warunku może być on czynnikiem wymaganym lub nieistotnym – analogicznie do powyższych definicji. Przykładowo przesłanką warunkującą dla niektórych modeli obliczania składki on-line jest przyjęcie przez powiązany czynnik określonej wartości - zadeklarowanie chęci ubezpieczenia dodatkowego ryzyka w ramach produktu może pociągnąć potrzebę dostarczenia dodatkowej informacji dla łączonego modelu wyceny pierwotnego i dodatkowego ryzyka.

Dla opisywanej metody ta subontologia stanowi narzędzie dodatkowego opisu zmiennych taryfikacyjnych oraz powiązanych z nimi pól formularza. W szczególności daje to możliwość wsparcia dla pól specjalnych, z takimi wartościami jak numery PESEL, VIN etc. W dalszej części subontologia daje też możliwość powiązania zmiennych taryfikacyjnych i niektórych ukrytych czynników ryzyka[217] (np. kod pocztowy a wypadkowość, gęstość zaludnienia czy liczba samochodów).

6.4 Metoda ekstrakcji modelu wyceny produktu ubezpieczeniowego

W podrozdziale 5.1 podano sposób rozumienia modelu wyceny na potrzeby niniejszej pracy oraz przeprowadzonych badań. Dokonano także uszczegółowienia pojęć, wprowadzając podział na model pierwotny (model wyceny składki) oraz model wtórny (model wyekstrahowany).

6.4.1 Założenia metody ekstrakcji modelu wyceny

Zadanie ekstrakcji modelu wyceny produktu ubezpieczeniowego zostało zdefiniowane oraz szczegółowo omówione w ramach podrozdziału 4.4. Zadanie to polega na odtworzeniu – albo bardziej precyzyjnie – zbudowaniu jak najbardziej zbliżonego modelu wtórnego na podstawie informacji o modelu pierwotnym oraz danych przez niego generowanych, a pozyskanych za pomocą źródła webowego. Oznacza to, że dla metody realizacji powyższego zadania:

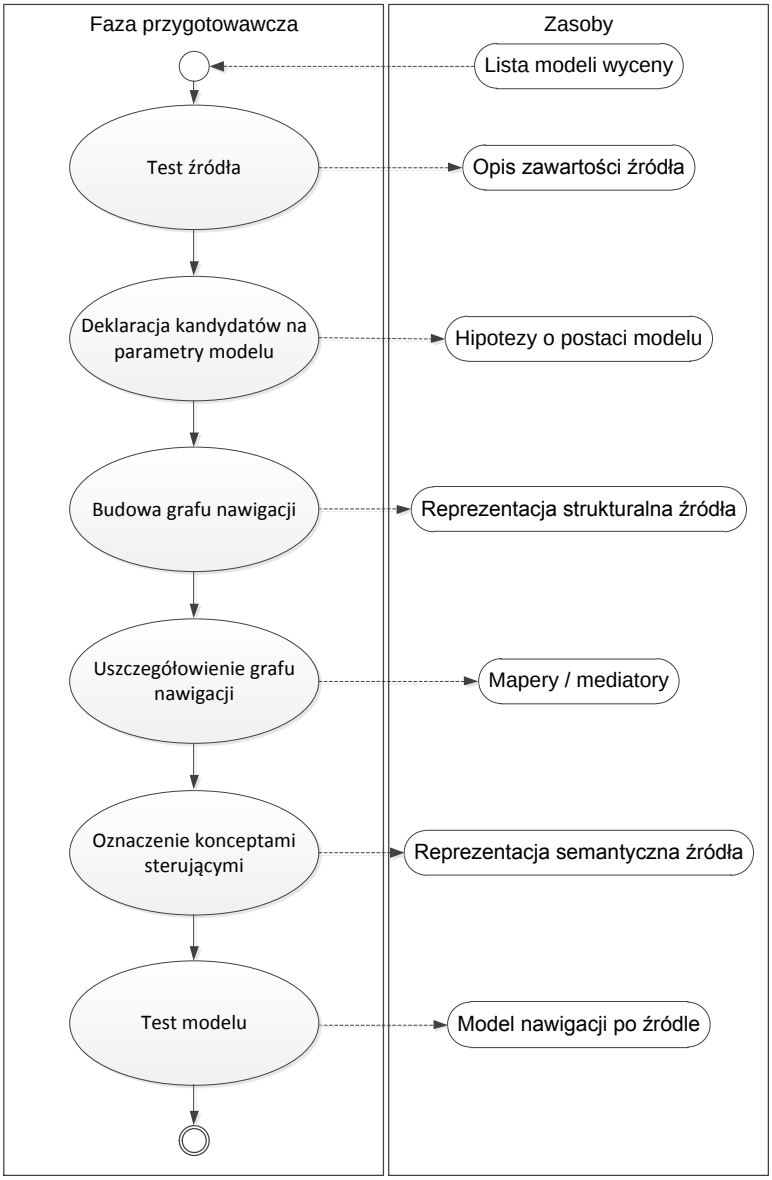
- wejściem: będzie źródło webowe, a dokładniej – dane i informacje związane z modelem pierwotnym (modelem wyceny składki),
- wyjście: stanowi model wtórny (czyli model wyekstrahowany) lub ewentualnie kilka konkurencyjnych modeli.

Ponadto w przypadku prezentowanej metody oraz uzyskanych za jej pomocą rezultatów, za dodatkowe wejście uznać można wiedzę przechowywaną w postaci ontologii, której kształt został zaproponowany w podrozdziałach 6.3.1-6.3.3.

Obecnie zajmiemy się przedstawieniem ogólnej metody pozwalającej na zrealizowanie zadania ekstrakcji modeli wyceny produktu ubezpieczeniowego ze źródła inter-

[217] Wiedza ta została wykorzystana przy tworzeniu wzbogaconych zbiorów danych w badaniu opisanym w rozdziale 7.

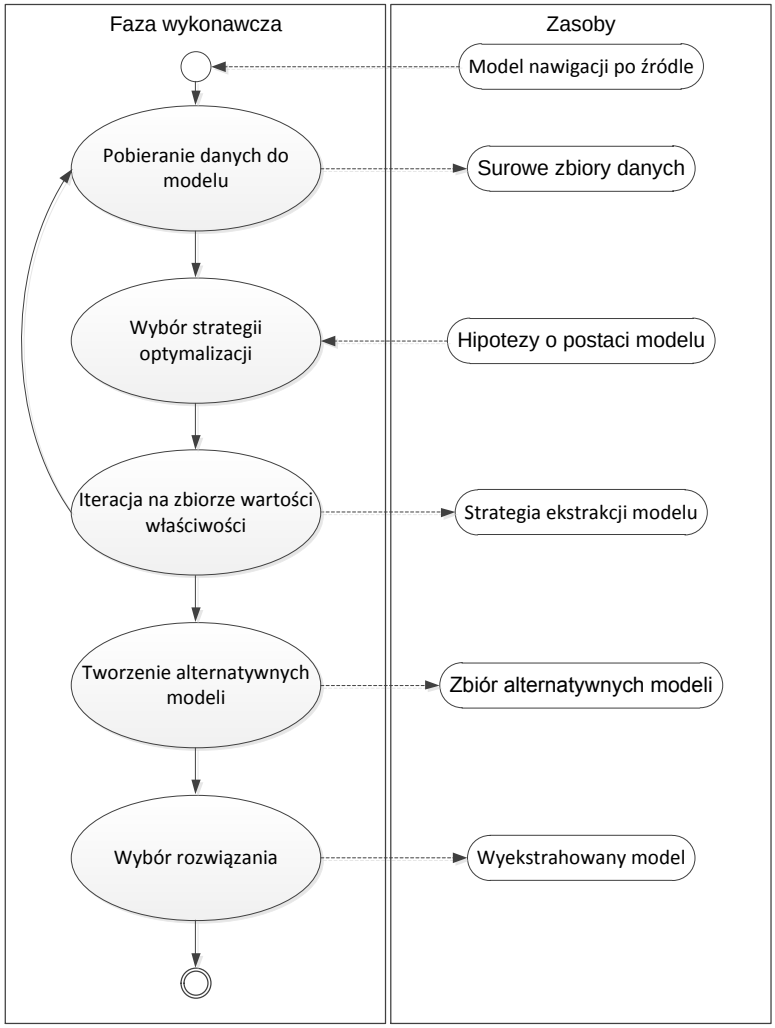
netowego. Generalny schemat idei proponowanej metody przedstawiony został na rysunkach 15 i 16.



Rysunek 15. Faza przygotowawcza procesu ekstrakcji modeli składki ze źródła webowego
Źródło: opracowanie własne na podstawie [Abramowicz2011]

Prezentacja metody podzielona została na fazy oraz etapy. W metodzie wyróżniono dwie fazy składające się z szeregu etapów. Faza pierwsza, to **faza przygotowawcza**. Obejmuje ona działania polegające na przeglądzie źródła oraz przygotowaniu artefaktów niezbędnych do pomyślnego wykonania dalszej części metody. Z kolei faza druga to **faza wykonawcza**, stanowiąca zasadniczą część realizacji metody. W ramach wyróżnionych w niej etapów odbywa się nawigacja po źródle internetowym. Źródło to jest odpytywane, a zwracane w ramach odpowiedzi na zapytania dane są gromadzo-

ne. Dalej dane te stanowią podstawę do zbudowania jednego lub większej liczby próbnych modeli wtórnych.



Rysunek 16. Faza wykonawcza procesu ekstrakcji modeli składki ze źródła webowego
Źródło: opracowanie własne na podstawie [Abramowicz2011]

Na wspomnianych już rysunkach przedstawiono diagramy przepływu odpowiednio dla fazy przygotowawczej (rysunek 15) oraz fazy wykonawczej (rysunek 16). Obydwa diagramy należy rozumieć jako całość, tzn. obydwie fazy przede wszystkim współdzielą zasoby i przewidziane są do uruchomienia w tym samym środowisku. Poza tym mogą być obsługiwane przez tego samego operatora, a druga faza może być zrealizowana niezwłocznie po przeprowadzeniu pierwszej z nich.

Diagramy przedstawiają przebieg procesów składających się na metodę z zaznaczonymi zgrupowanymi etapami[218]. Zgrupowania etapów połączone są ciągłymi strzałkami, których groty wskazują na kolejność realizacji poszczególnych elementów

[218] Stanowią one skrótowe uogólnienie poszczególnych etapów, których omówieniem zajmiemy się dalej.

przebiegu. Obok ścieżki przedstawiono pole przeznaczone do wskazania najistotniejszych elementów informacji oraz struktur wejścia-wyjścia użytych w ramach etapów. Elementy te wraz ze strukturami zbiorczo określać będziemy jako **artefakty metody**. Charakter wspomnianego wykorzystania reprezentowany jest przez kierunek przerywanych strzałek na diagramach. Strzałki skierowane od zgrupowanych etapów do artefaktu oznaczają wytworzenie lub zapełnienie struktury informacjami. Strzałki skierowane odwrotnie reprezentują przypadki wykorzystania danego artefaktu. Ważnym aspektem następstwa faz jest warunek możliwości realizacji fazy drugiej, polegający na prawidłowym i zakończonym sukcesem wykonaniu fazy przygotowawczej. Przez zakończenie sukcesem rozumieć należy pełne, poprawne i zgodne z wymogami metody wytworzenie wszystkich niezbędnych artefaktów. Wyróżnienie dwóch faz ma sens o tyle, że po jednokrotnym zrealizowaniu fazy przygotowawczej możliwe jest dla danego źródła i przy założeniu braku istotnych zmian w jego funkcjonowaniu[219] wielokrotne wykonanie fazy wykonawczej, a co za tym idzie, wielorazowe wygenerowanie modeli wtórnych. W praktyce taka sytuacja pozwala osiągnąć następujące korzyści: po pierwsze, możliwe jest rozłożenie w czasie realizacji fazy wykonawczej, o ile zajdzie taka potrzeba; po drugie, możliwe jest generowanie modeli w odstępach czasu, np. w celu weryfikacji ich zmienności.

Implementacja metody poza deklaracją procesu, zdefiniowaniem elementów i struktur gromadzenia i wymiany informacji oraz wiedzy w postaci wspomnianych artefaktów przewiduje także przygotowanie szeregu komponentów narzędziowych, których użycie spowoduje zwiększenie efektywności realizacji metody w praktyce oraz poprawy komfortu operatora metody na różnych etapach jego pracy. Narzędziami tymi są trzy komponenty:

1. Planista Nawigacji,
2. Eksplorator Reprezentacji Źródła,
3. Komponent Nawigacyjno-Ekstrakcyjny.

Szersze przedstawienie zaprojektowanych komponentów narzędziowych oraz poszczególne etapy zostaną szczegółowo omówione w podrozdziale 6.4.2.

[219] albo inaczej, braku konieczności aktualizacji jego formalnego opisu stworzonego w pierwszej fazie metody.

Jak już wcześniej zauważono, zadanie realizowane za pomocą przedstawianej metody w swoich założeniach jest zbliżone do ekstrakcji informacji ze źródeł głębokiego internetu. Istnieje jednak szereg ważnych różnic wymagających innego podejścia do tego pierwszego zagadnienia. Po pierwsze - w odróżnieniu od ekstrakcji informacji, w zaproponowanym podejściu ekstrahujemy wiedzę. Źródłem, do którego się odwołujemy, jest nie baza danych osiągalna przez interfejs webowy[220], lecz algorytm implementujący model wyceny[221]. Po drugie, liczba potencjalnych wyników jednostkowych w przypadku ekstrakcji modeli wyceny ze względu na zmianę kryteriów kalkulacji[222] może być zdecydowanie większa niż w przypadku ekstrakcji informacji z głębokiego internetu, z drugiej strony istotna część wyników może być potraktowana jako pomijalna bez spadku jakości modelu wtórnego[223]. Przedstawione w dalszej części badania z zastosowaniem opisanej metody wskazują, że możliwe jest wyekstrahowanie modelu z szerokiej klasy źródeł webowych[224]. przy uwzględnieniu pewnych założeń minimalnych[225]. Obecnie skoncentrujemy się na opisie poszczególnych etapów metody.

6.4.2 Etapy procesu ekstrakcji

Test źródła

Jest to pierwszy etap proponowanej metody. Jego celem jest zapoznanie operatora systemu ekstrahującego z potencjalnie nowym webowym źródłem wiedzy ubezpieczeniowej. Ważne jest, aby operator w tym kroku przeszedł przez pełen proces kalkulacji składki, zidentyfikował występowanie formularzy oraz pól warunkowych, a także możliwych sytuacji asynchronicznej wymiany danych z serwerem[226]. Dodatkowo ope-

[220] Najczęściej jest to pewien rodzaj formularza.
[221] Pozór podobieństwa powodować może fakt, iż wyniki działania badanego algorytmu poznajemy poprzez zbliżony do przypadku ekstrakcji z głębokiego internetu interfejs WWW.
[222] Ich analogiem w przypadku zadania ekstrakcji informacji z głębokiego Internetu są kryteria wyszukiwania w bazie danych
[223] W przypadku istnienia zależności funkcyjnych pomiędzy kryteriami a wysokością składki. Dla pewnych wartości zmiennej kryterium wynikowa wysokość składek może być wspólna lub niezmienna. Pojawia się jednak tutaj problem pogrupowania zmiennych kryterialnych oraz określenia przedziałów zmiennych kryterialnych.
[224] Przede wszystkim chodzi o witryny internetowe, ale potencjalnie także np. usługi sieciowe.
[225] Założenia, o których tutaj mowa dotyczą przede wszystkim technologicznych aspektów zwianych z ekstrakcją informacji z rozpatrywanego źródła, takich jak: kwestie możliwości uzyskania dostępu, intensywności dostępu do źródeł, stabilności otrzymywanych informacji, niewystępowania przeszkód w komunikacji etc.
[226] W warunkach korzystania z normalnego klienta WWW wykrycie wszystkich takich sytuacji może stanowić problem. Ułatwienie może stanowić tutaj zastosowanie oprogramowania przygotowanego dla celów realizacji metody, o którym mowa poniżej.

rator powinien zapoznać się ze sposobem prezentacji wyników. Operator dokonuje także w tym etapie wstępnej oceny technicznych aspektów wykonania źródła webowego pod kątem ewentualnych możliwych utrudnień w pracy silnika ekstrahującego. Zasadniczo test źródła wykonywany jest przez operatora z użyciem dowolnego klienta WWW, ale może być też wykorzystany Planista Nawigacji – komponent oprogramowania przygotowany na potrzeby kolejnych etapów realizacji metody.

Deklaracja kandydatów na parametry modelu

Ten etap oraz następny, w odróżnieniu od wcześniej wspomnianego testu źródła, realizowany jest za pomocą wsparcia ze strony specjalnie zaplanowanego dla metody komponentu oprogramowania. Dodatkowo opisywany etap oraz etap następny przewidziane są w zasadzie do wykonania w sposób równoległy. W ramach deklaracji kandydatów na parametry modelu operator dokonuje wskazania, które zmienne taryfikacyjne oraz dopuszczające do ubezpieczenia – albo zgodnie z nazewnictwem zaproponowanym w ontologii, jakie CzynnikiRyzyka – stanowią potencjalne wejście dla modelu wyceny składki i w związku z tym powinny być uwzględnione w ramach procesu ekstrakcji danych oraz przy tworzeniu modeli wyekstrahowanych.

Z punktu widzenia przytoczonej ontologii wskazanie to odbywa się przez wybór przez operatora istniejących ManifestacjiCzynnikówRyzyka. Technicznie oznacza to wskazanie oraz analizę określonych pól formularzy wypełnianych przez potencjalnego klienta w ramach procesu wyliczenia jego składki. Wykorzystywane w tym etapie narzędzie pozwala wszakże na zaznaczanie całych formularzy, dokonując następnie uproszczonej[227] analizy w sposób automatyczny.

Budowa grafu nawigacji

Planista Nawigacji pozwala na samoczynne tworzenie zaawansowanej wersji grafu nawigacji. Graf nawigacji stanowi najważniejszy element reprezentacji strukturalnej źródła webowego zgodnie z tym, jak zostało to przedstawione w podrozdziale 6.2. Przez zaawansowaną wersję grafu należy rozumieć graf nawigacji w takim stanie, który może wymagać dodatkowych ingerencji lub modyfikacji ze strony operatora w celu

[227] Ewentualne błędy oprogramowania można skorygować na późniejszych etapach, m.in. za pomocą oznaczenia zidentyfikowanych pól konceptami sterującymi.

otrzymania funkcjonalnej wersji. Konieczność oraz intensywność tych ingerencji zależna jest od stopnia skomplikowania reprezentowanego źródła internetowego.

Samoczynność budowy grafu przez Planistę Nawigacji osadza się na generowaniu obiektowej[228] wersji grafu stworzonego na podstawie odwzorowania pojedynczego[229] przebiegu nawigacji po źródle wykonanego przez operatora z wykorzystaniem komponentu narzędziowego, który rejestruje wszystkie parametry operacji użytkownika oraz przypadki komunikacji klienta z serwerem webowym. Równocześnie z odbywającą się rejestracją przebiegu operator ma możliwość wyboru oraz wskazania szeregu elementów i czynników istotnych z punktu widzenia późniejszej ekstrakcji danych oraz uogólnienia przypadku nawigacji[230]. Na podstawie wszystkich wymienionych informacji algorytmy Planisty Nawigacji wykorzystują podstawowe heurystyki dla generalizacji przykładu nawigacji. Automat sprawdza także możliwe alternatywne operacje związane np. z obsługą formularzy w stosunku do wskazanych przez operatora[231].

Planista Nawigacji jest komponentem programowym oferującym operatorowi zaawansowany interfejs użytkownika. Daje on wsparcie przy pracy nad kilkoma etapami fazy przygotowawczej: omówionej powyżej deklaracji kandydatów na parametry modelu, referowanego obecnie etapu budowy grafu nawigacji, a także oznaczenia konceptami sterującymi.

Uszczegółowienie grafu nawigacji

Zgodnie z opisem poprzedniego etapu metody domniemywa się, że wygenerowane przez komponent planowania nawigacji artefakty nie stanowią w pełni użytecznego materiału do wykorzystania w fazie wykonawczej metody. Uszczegółowienie grafu nawigacji stanowi etap pośredni mający na celu usunięcie wszelkich nieprawidłowości w dotychczas przygotowanej wersji reprezentacji źródła, uzupełnienie brakujących elementów[232] oraz skorygowanie ewentualnie niepoprawnie funkcjonujących heury-

[228] Zapis w XML następuje na końcu, w ramach procesu serializacji uzyskanych obiektów.
[229] Zastosowane rozwiązanie nie ma charakteru uczenia maszynowego z wielu przykładów.
[230] m.in. deklaracja ekstraktorów, dynamicznych formularzy, przypadków asynchronicznej komunikacji z serwerem.
[231] W ten sposób generowana jest przestrzeń warunkowych zbiorów wartości właściwości.
[232] np., o uszczegółowienie polityki związanej z nienarzucaniem się ze zbyt agresywnym odpytywaniem serwera oraz obsługę wariantów przeciwdziałania zabezpieczeniom serwera poprzez obsługę mechanizmów proxy lub CAPTCHA.

styk zmierzających do uogólnienia grafu nawigacji na dostatecznie szeroki zakres przypadków nawigacji po analizowanym źródle.

Większość z przewidzianych na ten etap działań operator wykonać może przy użyciu narzędzi dostarczonych wraz z metodą. Pozostała część omawianych operacji wymaga bezpośredniej ingerencji w XML-owy opis strukturalny i semantyczny źródła[233].

Oznaczenie pojęciami sterującymi

Zgodnie z zaproponowanym sposobem modelowania oraz opisu źródła webowego, każdy element reprezentacji strukturalnej źródła może być dodatkowo opatrzony zbiorem oznaczeń (adnotacji) wykorzystujących pojęcia z ontologii oraz poszczególnych jej części, zaprezentowanej w podrozdziałach 6.3 i następnych.

Jak zauważono na poziomie ogólnego opisu metody w podrozdziałach 5.2-5.3, oznaczenia te, a w szczególności niesione przez pojęcia znaczenie, wykorzystane są przez silnik ekstrakcji realizujący swoje zadania w fazie wykonania, do racjonalnego postępowania oraz efektywniejszego wykorzystania zasobów udostępnianych przez serwer webowy.

Adnotowanie pojęciami wykonane może być przez operatora na etapie wykorzystania Planisty Nawigacji. Komponent ten przez prezentowany interfejs użytkownika pozwala w określonych cyklach pracy na włączanie mechanizmu oznaczania pojęciami oraz udostępnia możliwość wyboru konkretnego terminu z ontologii. W rezultacie użytkownik może w formie wizualnej interakcji wskazać pojęcie oraz element, który ma zostać z danym pojęciem skojarzony.

Informacja o przypisaniu pojęć jest pierwotnie przechowywana w źródle (X)HTML-owym dokumentów pochodzących ze źródła internetowego. Następnie jest przenoszona na odpowiadające poszczególnym znacznikom (X)HTML elementy opisu strukturalnego źródła zapisanego w postaci dokumentu XML.

Test modelu

Etap testowania modelu polega na syntaktycznej oraz semantycznej walidacji dokumentu zawierającego reprezentację źródła webowego. Zasadniczo odbywa się

[233] Nie stanowi to o ułomności metody a jedynie o prototypowym charakterze komponentów narzędziowych, które wymagałyby poświecenia większej ilości czasu na dopracowanie i implementację dodatkowych funkcji.

to na kilka sposobów. Weryfikacja na poziomie syntaktycznym może być wykonana przez dowolne narzędzie posiadające możliwość parsowania dokumentu XML w kontekście jego poprawności ze standardem tego języka oraz przygotowanym w ramach prezentowanej metody schematu właściwego dokumentu XML (XML Schema[234]). Druga możliwość testu na tym poziomie polega na wczytaniu dokumentu do prototypowego środowiska integrującego wspomniane w podrozdziale 6.4 komponenty narzędziowe. Dokument z reprezentacją źródła internetowego jest wczytywany przez Eksplorator Reprezentacji Źródła. Podczas tego procesu plik podlega weryfikacji, a wszelkie nieprawidłowości w jego konstrukcji są wskazywane operatorowi.

Z kolei walidacja semantyczna dokumentu może polegać na eksperckiej ocenie dokonanej przez operatora systemu lub też na testowym wykonaniu pojedynczego przebiegu ekstrakcji przez Komponent Nawigacyjno-ekstrakcyjny w oparciu o testowany opis źródła. Przegląd dowolnego dokumentu zawierającego reprezentację źródła webowego jest istotnie ułatwiony dzięki wspomnianemu już Eksploratorowi Reprezentacji Źródła.

Eksplorator Reprezentacji Źródła jest komponentem narzędziowym służącym do graficznej prezentacji grafu nawigacji oraz wygodnego analizowania wszystkich struktur informacji oraz obiektów wchodzących w skład dowolnej reprezentacji źródła. Struktury te oraz obiekty zostały wymienione w podrozdziałach 5.3-5.4 oraz 6.2. Poza prezentacją narzędzie to daje także ograniczone możliwości wizualnego edytowania[235] załadowanych do środowiska[236] dokumentów z reprezentacjami.

Pobieranie danych do modelu

Etap pobierania danych do modelu realizowany jest w pełni automatycznie przez Komponent Nawigacyjno-Ekstrakcyjny. Równolegle do tego etapu realizowany jest kolejny krok metody, tj. podejmowanie decyzji odnośnie strategii optymalizacji.

Komponent Nawigacyjno-Ekstrakcyjny jest automatem, którego zadaniem jest cykliczne wykonywanie procesu nawigacji po źródle webowym, zgodnie z planem wynikającym z dostarczonej dokumentacji źródła internetowego. Komponent, realizując

[234] XML Schema jest językiem umożliwiającym tworzenie opisów pozwalających zweryfikować poprawność (walidować) odpowiadające tym opisom dokumenty XML.
[235] Proste czynności polegające przede wszystkim na usuwaniu i dodawaniu wierzchołków określonych typów.
[236] Środowisko zaprojektowane zostało do funkcjonowania w trybie wielodokumentowym.

swoje zadanie, iteruje w ramach każdego cyklu nawigacji przez działanie operatora inkrementacji na przestrzeni wartości właściwości.

```
operator ⊕
begin
  W^n := W \ W^w
  i := |W^n|
  iteruj := true
  foreach (w ∈ W^n) do
  begin
    A := adnotacje(w)
    if (A ∪ A^w ≠ Ø)
    begin
      if (czy pusta(wartość(w)))
      begin
        reset(w)
        zapisz stan właściwości(w)
        i--
        continue
      end
      else
      begin
        if (not(zbiór wartości(w) = Ø))
        begin
          if (not(iteruj))
          w := następny(w)
        end
      end
    end
    if (czy losowa(w))
      w := następny(w)
    else
    begin
      if (iteruj)
        w := następny(w)
    end
    zapisz stan właściwości(w)
    if (not(przekroczono zakres(w))
      iteruj := false
    else
      i--
  end
  return (i <= 0)
end
```

Rysunek 17. Algorytm iteratora na zbiorze wartości właściwości
Źródło: opracowanie własne

Iteracja na zbiorze wartości właściwości ma charakter globalny i działa na zasadzie hierarchicznej, analogicznie do zasady działania licznika. Realizacja procesu iterowania wykonana jest za pośrednictwem operatora, którego funkcjonowanie opisuje algorytm zaprezentowany na rysunku 17. Sam operator nie uwzględnia heurystyk optymalizacji opisanych dalej dzięki temu jest on niezależny od ich zastosowania oraz szczegółowej implementacji.

Warunkiem stopu automatu wykonującego proces nawigacji jest warunek, którego spełnienie powoduje nierozpoczynanie kolejnego cyklu nawigacji po źródle. W przypadku grafu opisującego proces nawigacji w prototypowym rozwiązaniu warunek taki może być określony jako zrealizowanie zadanej liczby cykli nawigacji po gra-

fie lub realizacja cykli nawigacji powodująca wyczerpanie możliwości dalszego iterowania na przestrzeni wartości właściwości.

Ograniczenie liczby cykli nawigacji po grafie nie ma bezpośredniego wpływu na działanie operatora iteracji. Ma natomiast wpływ na zachowanie heurystyk opisanych dalej.

Wybór strategii optymalizacji

Podczas zbierania danych ze źródła wiedzy ubezpieczeniowej eksponującego model wyceny składki wykonywany jest jednocześnie proces wyboru heurystyk umożliwiających zmniejszenie liczby cykli nawigacji.

Dążenie zaproponowanej metody do redukcji cykli nawigacji po źródle webowym, a co za tym idzie, zmniejszenie liczby zapytań do serwera, stanowi ważny aspekt praktycznej realizacji rozwiązania problemu. Jest to także zgodne z jego opisem zawartym w podrozdziale 4.4 oraz rozdziale 1. Próby optymalizacji dokonane w przedstawionym zakresie są niezbędne ze względu na czas generowania modelu, jak też na możliwe ograniczenia w komunikacji z serwerem, mogące stanowić wynik działania różnorodnych mechanizmów. Tabela 10 prezentuje heurystyki możliwe do wykorzystania w celu redukcji cykli nawigacji wraz z ich głównymi założeniami.

W prototypowym rozwiązaniu wykorzystanym dla celów badawczych zaimplementowano wszystkie z przedstawionych strategii z wyjątkiem strategii 7. Bardziej szczegółowe omówienie heurystyk z nimi powiązanych znajduje się w podrozdziałach 6.4.3 i 6.4.4.

Tworzenie alternatywnych modeli

Przedostatni etap stanowi wykorzystanie artefaktów powstałych w wyniku realizacji poprzednich kroków metody dla opracowania jednego lub większej liczby alternatywnych modeli wtórnych. Szczególnym zasobem wykorzystanym do tego zadania są dane gromadzone i przechowywane w wyniku pracy Komponentu Nawigacyjno-Ekstrakcyjnego.

W przytaczanym już kilkukrotnie badaniu wstępnym zdecydowano się na generowanie modeli wtórnych za pomocą trzech konkurencyjnych metod eksploracji danych. W przeprowadzonym na potrzeby niniejszej pracy badaniu zakres testowanych narzędzi analitycznych został znacznie rozszerzony. Podejście polegające na zastosowaniu

wielu konkurencyjnych narzędzi analitycznych ma szereg zalet. Po pierwsze, pozwala zbadać, która z technik nadaje się najlepiej do przetwarzania danych i wiedzy przy uwzględnieniu specyfiki metody oraz dziedziny. Po drugie, poza oczekiwanym zróżnicowaniem w jakości uzyskanych alternatywnych modeli, poszczególne techniki analityczne mają różne założenia co do postaci samego modelu. Wreszcie posiadanie konkurencyjnych modeli daje możliwość w zakresie ich wzajemnego porównywania oraz weryfikacji.

Tabela 10. Podejścia związane z wyborem strategii optymalizacji liczby zapytań dla budowy modelu
Źródło: opracowanie własne

L.p.	Heurystyka	Uwagi	Implementacja
1	Brak optymalizacji – podejście naiwne.	Nieefektywne	-
2	Naiwne odkrywanie liniowości przy założeniu ceteris paribus.	Heurystyka prosta w implementacji, dająca zadowalające rezultaty	Pełna
3	Naiwne odkrywanie braku wpływu parametru polisy.	Podstawowa heurystyka. Ważna do wykrycia zmiennych nieistotnych	Pełna
4	Posiadanie dodatkowej wiedzy o parametrze polisy.	Wiedza przechowywana w adnotacjach właściwości. Punkt odniesienia – teoria zebrana w podrozdziale 3.2	Pełna
5	Posiadanie dodatkowej wiedzy o modelu.	Wiedza przechowywana w adnotacjach modelu źródła. Punkt odniesienia – analiza innych wyekstrahowanych modeli	Pełna
6	Odkrywanie zależności funkcyjnych lub braku wpływu poprzez wnioskowanie statystyczne przy założeniu ceteris paribus.	Implementacja wykrywania prostych zależności funkcyjnych	Częściowa
7	Odkrywanie zależności funkcyjnych lub braku wpływu poprzez wnioskowanie statystyczne z uwzględnieniem wpływu zmian innych zmiennych niezależnych modelu.	Skomplikowana implementacja. Niski oczekiwany zysk optymalizacyjny w stosunku do nakładów implementacyjnych	Brak

Wprowadzenie do tematyki eksploracji danych przedstawiono w podrozdziale 2.4-2.4.4. W omawianym zestawie eksperymentów, jako aparatu badawczego, użyto technik wyszczególnionych w przytoczonym wprowadzeniu do tematyki eksploracji danych, a także szeregu innych technik.

W efekcie zastosowania powyżej wymienionych metod analitycznych na danych pozyskanych w wyniku wykonania dwóch poprzednich etapów możliwe jest stworzenie alternatywnych modeli wtórnych dla wyceny konkretnego produktu ubezpieczeniowego.

Wybór rozwiązania

Wybór konkretnego modelu wtórnego spośród zaproponowanych w trakcie wykonania przedostatniego kroku metody konkurencyjnych modeli wynikających z zastosowania różnorodnych technik analizy danych jest etapem kończącym zastosowanie prezentowanej metody. Wyboru tego dokonuje operator systemu stanowiącego implementację metody.

Tabela 11. Proces ekstrakcji modeli wyceny produktu ubezpieczeniowego ze źródła internetowego
Źródło: opracowanie własne

Faza	Etap	Automatyzacja	Wsparcie	Uwagi
Przygotowanie				
	Test źródła	Brak	Opcjonalne	
	Deklaracja kandydatów na parametry modelu	Pełna	Przez interfejs użytkownika	
	Budowa grafu nawigacji	Pełna	Przez interfejs użytkownika	
	Uszczegółowienie grafu nawigacji	Częściowa	-	Edycja XML
	Oznaczenie konceptami sterującymi	Brak	Przez komponent Planisty Nawigacji	Etap opcjonalny
	Test modelu	Pełna	Przez moduł ekstraktora	
Wykonanie				
	Wybór strategii optymalizacji	Częściowa	-	
	Pobieranie danych do modelu	Pełna	Przez moduł ekstraktora	
	Tworzenie alternatywnych modeli	Częściowa	Oprogramowanie analityczne	
	Wybór rozwiązania	Brak	-	

Można wskazać na pewną liczbę czynników stanowiących przesłanki dla racjonalnego procesu decyzyjnego w omawianym zakresie. Są nimi w szczególności[237]:

- oczekiwana jakość modeli rozumiana jako dopasowanie modelu wtórnego do pierwotnego,
- wyniki ewaluacji modeli z innych źródeł,
- pożądane cechy wyniku konkretnej techniki eksploracji danych,
- zastosowanie wybranego modelu wtórnego,
- charakter źródła wiedzy ubezpieczeniowej.

W tabeli 11 dokonano zestawienia oraz podsumowania opisanych etapów metody. Zestawienie uzupełniono dodatkowo o uwagi oraz wyszczególnienie cech charakterystycznych poszczególnych etapów.

6.4.3 Modele ze stanami dyskretnymi

Szacunkowa liczba wszystkich możliwych wartości parametrów testowych modeli w przeprowadzonym badaniu wahała się w granicach[238] od 144000 do $8{,}6*10^9$. Taka liczba zapytań wyklucza zastosowanie rozwiązania polegającego na braku optymalizacji (strategia 1 w tabeli 10).

Strategie 4 i 5 w eksperymentalnym prototypie zostały zrealizowane przez wprowadzenie możliwości ręcznego przypisania pojęć z ontologii dziedzinowej do poszczególnych elementów modelu źródła internetowego. Dodatkowa wiedza daje możliwość usprawnienia działania systemu ekstrahującego, np. poprzez przypisanie predefiniowanego zbioru wartości dla określonego parametru polisy.

W przypadku zmiennych dyskretnych wyznaczanie zależności funkcyjnych w postaci formuł matematycznych może być niewykonalne. Jeżeli zmienna posiada niewielki zbiór wartości (np. zmienne binarne), to możliwe jest wytypowanie prawdopodobnych innych zmiennych, z którymi zmienna dyskretna może wchodzić w interakcję generując różne wyniki końcowe modelu. W przypadku zmiennych dyskretnych posiadających większy zbiór wartości zastosowanie ma z kolei losowy dobór wartości dla tej zmiennej.

[237] Szczegółowe kwestie dotyczące oceny modeli uzyskanych w badaniu zostały zaprezentowane w podrozdziałach 7.3-7.5.

[238] Podane wielkości mają charakter poglądowy. Wyznaczenie mocy zbioru wszystkich wartości zależy od dziedziny manifestacji czynników ryzyka, ale także różnego rodzaju założeń.

6.4.4 Modele liniowe i nieliniowe

Przez pojawiające się w strategiach 2 oraz 6 założenie ceteris paribus rozumiemy sytuację, w której badamy zachowanie zmiennej zależnej tylko i wyłącznie ze względu na zmianę jednej zmiennej niezależnej.

W praktyce lepszym podejściem okazuje się podział zbioru wszystkich zmiennych. Podział poprzedzony jest wytypowaniem zmiennych, które prawdopodobnie nie mają wpływu na wyniki modelu, a następnie podział pozostałych zmiennych na względnie niezależne podzbiory[239]. Następnie zapytania wykonywane są poprzez iterację nie na całym zbiorze zmiennych tylko poprzez iterację na poszczególnych podzbiorach.

Strategia 6 umożliwia badanie dwóch rodzajów hipotez: o braku wpływu wybranego parametru na model przy określonym poziomie istotności albo o znanej postaci funkcyjnej zależności tego parametru i zmiennej zależnej przy określonym poziomie istotności.

Zmienne ciągłe są naturalnymi kandydatami na zmienne, dla których można próbować znaleźć funkcję matematyczną określającą zależność między taką zmienną i wynikiem modelu. W szczególności zależność liniowa jest oczekiwana od zmiennej, która pełni rolę miary ekspozycji na ryzyko[240]. W takim przypadku istnieje możliwość znacznej redukcji zapytań, a w dalszej części procesu ekstrakcji taką zmienną można w zasadzie pominąć.

6.5 Prototypowa implementacja

Elementem opisanej w niniejszej pracy metody jest prototypowe rozwiązanie stanowiące system integrujący komponenty narzędziowe służące do wsparcia w realizacji empirycznego badania. Badanie to ma za zadanie stanowić potwierdzenie koncepcji oraz zapewnić materiał niezbędny do dokonania ewaluacji metody oraz wytworzonych w związku z jej zastosowaniem artefaktów.

[239] Taki podział może wynikać z wiedzy zewnętrznej, już zebranego wstępnie zbioru wyników generowanych przez model lub może mieć charakter genetyczny (zmienne opisujące pokrewny element ryzyka).
[240] Por., podrozdział 3.2.2. Identyfikacja takich zmiennych jest w ograniczonym zakresie możliwa na podstawie zgromadzonej wiedzy.

Zgodnie z informacją zawartą w podrozdziale 6.4 zrealizowane środowisko składa się z trzech komponentów narzędziowych. W opracowaniu prototypu istotną rolę odegrało wykorzystanie szeregu modułów oraz niezależnych komponentów programowych, których kod został wykorzystany lub zintegrowany w rozwijanym systemie. Samo prototypowe środowisko zostało wykonane w nowoczesnej technologii aplikacji desktopowej wykorzystującej ramę WPF[241]. Środowisko napisane zostało w języku C#[242] oraz pomocniczo Java[243]. W związku z zastosowaną technologią oraz z powodu zależności niektórych użytych komponentów programowych prototyp do uruchomienia wymaga środowiska systemowego Windows. Wymóg taki podyktowany jest przede wszystkim wykorzystaniem udostępnionego w formie otwartego oprogramowania modułu cExWB[244]. Moduł ten stanowi opakowanie za pośrednictwem technologii COM interfejsu udostępnionego przez przeglądarkę internetową Internet Explorer. Ze względu na pośrednie wykorzystanie mechanizmów jej funkcjonowania zainstalowanie tego klienta WWW w systemie, na którym wykonywany jest prototyp, jest konieczne[245].

Drugim niezależnym komponentem programowym wykorzystanym przez zaprojektowane oprogramowanie jest biblioteka Graph#[246]. Jest to biblioteka dostarczająca kompletne API dla obsługi abstrakcyjnych grafów. Biblioteka ta daje także możliwości wizualizacji takich grafów.

Kolejnym modułem użytym w prototypie jest rozszerzenie HtmlXPath. Jest to kolejny otwarty projekt pozwalający na wykorzystanie standardowego API obsługi zapytań języka XPath w kontekście dowolnego dokumentu HTML, przy czym nie ma w tym wypadku znaczenia, czy dokument ten spełnia wytyczne dla poprawności standardu XHTML.

Dla obsługi bazy wiedzy wykorzystano otwartą bibliotekę OwlAPI[247]. Biblioteka ta działa w środowisku uruchomieniowym maszyny wirtualnej Java. Zastosowanie

[241] Windows Presentation Framework. http://msdn.microsoft.com/en-us//library/ms754130.aspx, odczytano 15-05-2015 r.
[242] http://www.ecma-international.org/publications/files/ECMA-ST/Ecma-334.pdf, odczytano 15-05-2015 r.
[243] http://www.java.com/, odczytano 15-05-2015 r.
[244] http://code.google.com/p/csexwb2/, odczytano 15-05-2015 r.
[245] Rozwiązanie przeznaczone jest tylko dla komputerów działających pod kontrolą systemów z rodziny Windows.
[246] http://graphsharp.codeplex.com/, odczytano 15-01-2014 r.
[247] http://owlapi.sourceforge.net/, odczytano 15-01-2014 r.

mechanizmu wnioskowania na ontologii dodatkowo wymagało użycia automatu wnioskującego Pellet[248]. Biblioteka OwlAPI ma standardową możliwość połączenia jej z interfejsem tego automatu. Pellet jest również natywnym rozwiązaniem działającym na maszynie wirtualnej Java.

W celu połączenia i integracji środowisk maszyn wirtualnych Java oraz CLI[249] użyto mostu IKVM[250].

Prace w ramach rozwoju prototypu objęły następujące działania:

1. integracja poszczególnych komponentów i modułów,
2. stworzenie GUI,
3. zaprojektowanie i zaimplementowanie w całości silnika dla Komponentu Nawigacyjno-ekstrakcyjnego,
4. oprogramowanie logiki aplikacji oraz obsługi API wymienionych komponentów i modułów dla realizacji funkcjonalności Planista Nawigacji i Eksplorator Reprezentacji Źródła,
5. zaprogramowanie użycia aplikacji zewnętrznych.

Tabela 12. Statystyki opisujące implementację rozwiązania
Źródło: Microsoft Visual Studio

Miara[251]	Wartość
Liczba projektów	10
Indeks utrzymania kodu	85
Złożoność cyklomatyczna	4163
Głębokość dziedziczenia	10
Sprzężenie klas	300
Liczba linii kodu	39526

Etap tworzenia alternatywnych modeli zrealizowany został poprzez użycie dodatkowych aplikacji. Aplikacjami tymi są specjalistyczne narzędzia służące do realizacji zadań eksploracji danych. Dla wszystkich metod eksploracji z wyjątkiem programowania genetycznego wykorzystane zostało oprogramowanie w ramach systemu SAS[252]

[248] http://clarkparsia.com/pellet/, odczytano 15-01-2014 r.
[249] ang. Common Language Infrastructure.
[250] http://www.ikvm.net/, odczytano 15-01-2015 r.
[251] Opisy podanych miar można znaleźć na stronie https://msdn.microsoft.com/pl-pl/library/bb385914.aspx, odczytano 31-05-2015 r.
[252] Wersja 9.4. Informacje o produktu: http://support.sas.com/software/94/index.html, odczytano 31-05-2015 r.

wraz z rozszerzeniem Enterprise Miner[253]. Natomiast modele generowane za pomocą metody programowania genetycznego zostały wytworzone przez odrębne oprogramowanie[254] (system SAS nie wspiera tego narzędzia analitycznego).

Cały prototypowy system zaprogramowany został z wykorzystaniem środowiska deweloperskiego Visual Studio. Tabela 12 zawiera najistotniejsze informacje charakteryzujące i syntetycznie podsumowujące prace dotyczące implementacji rozwiązania.

[253] Wersja 13.2. Karta produktu: http://support.sas.com/software/products/miner/, odczytano 31-05-2015 r.
[254] Discipulus 5.2 firmy RML Technologies, Inc.

7 Metodyka ewaluacji i ocena rozwiązania

7.1 Pozyskanie i analiza materiału badawczego

Materiał badawczy został pozyskany w ramach prowadzonych prac eksperymentalnych w okresie pomiędzy grudniem 2013 roku a marcem 2015 roku. W tym czasie przeprowadzono proces ekstrakcji na 19 źródłach danych ubezpieczeniowych (19 witryn WWW), należących do 13 firm ubezpieczeniowych. Dobór witryn przeprowadzono w oparciu o wytyczne opisane w podrozdziale 6.1. Cały proces eksperymentu wykonano zgodnie z metodą opisaną w rozdziale 6 oraz z wykorzystaniem modeli zaproponowanych w rozdziale 5 i podrozdziale 6.3.

Dla procesu badawczego rozwinięta została skomplikowana infrastruktura w celu pokonania szeregu wyzwań technicznych. Przede wszystkim, jeśli chodzi o ubezpieczenia motoryzacyjne, to stworzono ramę w języku JavaScript uruchamianą w konsoli przeglądarki internetowej w celu pozyskania wpisów w słowniku pojazdów poszczególnych ubezpieczycieli w ubezpieczeniach motoryzacyjnych (dla każdego ubezpieczyciela pobierane były kwotowania ubezpieczeń dla takich samych lub analogicznych modeli, typów oraz roczników pojazdów). Słowniki te zostały włączone do ontologii.

W wyniku przeprowadzonych prac zebrano dokładnie 40108 rekordów surowych danych. Każdy rekord surowych danych jest wynikiem pojedynczego pełnego cyklu nawigacji po źródle webowym. Rekord taki, może zawierać więcej niż jedną wartość składki. Zarówno spis wszystkich wykorzystanych źródeł oraz podział przypadającej liczby rekordów surowych danych na konkretne źródło danych znajduje się w podrozdziale 7.5.

Jeśli chodzi o statystyki czasowe poszczególnych etapów prowadzonych prac badawczych, to średni czas budowy pojedynczego grafu nawigacji za pomocą modułu Planista Nawigacji wyniósł 7 minut 48 sekund. Przy czym należy zauważyć, że w ostatecznym rozrachunku utworzone w ten sposób grafy nawigacji wymagały intensywnych modyfikacji o charakterze manualnym. Czas tych modyfikacji wielokrotnie przekraczał zmierzony średni czas budowy grafu.

Przeciętny czas adnotowania pojedynczego grafu nawigacji, w wariancie podstawowym, wyniósł 43 minuty i 49 sekund. Zmierzony czas odpowiadał oznaczeniu ele-

mentów reprezentacji semantycznej źródła przy okazji konstrukcji grafu nawigacji. W praktyce adnotacje te były wielokrotnie poprawiane i modyfikowane w dalszej części procesu ekstrakcji.

Wreszcie średni czas pozyskania pojedynczego rekordu (a więc także cyklu nawigacji po źródle webowym) wyniósł 3 minuty 17 sekund[255]. Odchylenie standardowe od tego czasu wyniosło 55 sekund. Wyliczenie to wyłącza cykle niekompletne oraz zakończone błędami.

W dalszym ciągu eksperymentu, dla każdego zbioru danych zrealizowano taką samą procedurę badawczą, która polegała na wykonaniu następujących czynności:

1. oczyszczeniu danych,
2. uzupełnieniu wartości pustych,
3. uzgodnieniu dat i normalizacja wartości,
4. konwersji formatów.

Tak przetworzone dane stanowiły przedmiot dalszego przetwarzania:

- kwalifikacji i podziału rekordów złożonych (pierwszy wariant),
- wzbogaceniu zbiorów danych o dane zewnętrzne (drugi wariant).

Podział rekordów złożonych (pierwszy wariant) odnosi się do rekordów uzyskanych z tych źródeł, które w ramach pojedynczego kwotowania zwracały więcej niż jedną obliczoną wartość (kilka wersji tego samego ubezpieczenia lub wartość składki dwóch odrębnych produktów, np. OC i AC pojazdów mechanicznych). W takim przypadku rekordy surowych danych zostały bądź zduplikowane, tyle że z pojedynczą wartością zmiennej zależnej (celu) lub wprowadzony został podział na oddzielne zbiory danych. W rezultacie tego podziału otrzymano 173466 rekordów z pojedynczą wartością składki.

W ramach eksperymentu podjęto także decyzję o próbie zweryfikowania hipotezy, że dodanie dodatkowych parametrów (zmiennych) do rekordów danych może wpłynąć na jakość otrzymanych modeli. W tym celu dokonano wzbogacenia poszczególnych zbiorów danych uzyskanych po dokonaniu podziału w pierwszym wariancie o dodatkowe zmienne nie należące do parametrów podawanych w ramach źródła WWW.

[255] Zmierzona wartość uwzględnia tzw. „politeness policy", tzn. odczekanie pewnego czasu między poszczególnymi połączeniami z serwerem, co jest normalną praktyką w tego typu przypadkach.

Wzbogacenie nastąpiło w przypadku występowania w danym zbiorze danych parametrów o charakterze porządkowym lub jakościowym, co do których istniało silne podejrzenie, że będą one miały istotny wpływ na kształtowanie się składki. Warunkiem wzbogacenia była dostępność odpowiadających danych pozwalających opisać parametry o charakterze porządkowym lub jakościowym w sposób ilościowy. Przykładowo dla ubezpieczeń motoryzacyjnych, w których występowała zależność od rodzaju pojazdu lub regionu geograficznego wśród danych wzbogacających zamieszczono m.in. dane demograficzne związane z regionem, takie jak gęstość zaludnienia, liczba ludności, dane dotyczące transportu – liczba pojazdów zarejestrowanych na obszarze oraz dodatkowe dane dotyczące pojazdu – wycena, typ nadwozia, pojemność, moc, liczba drzwi etc. Dane dotyczące pojazdu były uzupełniane wówczas, gdy dany kalkulator nie wymagał lub nie przewidywał identyfikacji pojazdu za pomocą określonej zmiennej.

Tabela 13. Zestawienie wszystkich zbiorów danych oraz liczebności ich rekordów
Źródło: opracowanie własne

L.p.	Zestaw danych po podziałach	Łączna liczba rekordów
	Niemotoryzacyjne	
1	aviva1A	4257
2	aviva1B	4257
3	ehome.benefia241A	22750
4	kuke1A.com	5445
5	kuke1B.com	2723
6	signal-iduna1A	1202
7	signal-iduna1B	1202
8	skokubezpieczenia24.home1A	7765
9	skokubezpieczenia24.home2A	1842
10	skokubezpieczenia24.health1A	5451
11	tutum.bike1A	3455
12	tutum.nnw1A	3277
13	uniqa241A	9576
14	uniqa242A	14359
15	youcandrive.home1A	6160
16	youcandrive.home1B	6160
17	youcandrive.travel1A	8077
18	youcandrive.travel1B	8077
19	youcandrive.travel2A	8077
20	youcandrive.travel2B	8077

Motoryzacyjne		
21	allianz1A	3260
22	allianz1B	3260
23	allianz2A	3260
24	allianz2B	3260
25	axadirect1A	2382
26	axadirect1B	2382
27	axadirect2A	4763
28	axadirect2B	4763
29	emoto.benefia241A	2393
30	emoto.benefia241B	2393
31	libertydirect1A	7603
32	libertydirect1B	7603
33	libertydirect2A	2607
34	libertydirect2B	2607
35	link41A	5943
36	link41B	5943
37	link42A	11483
38	link42B	11483
39	mtusa1A	2704
40	mtusa1B	2704
41	skokubezpieczenia24.moto1A	3879
42	skokubezpieczenia24.moto1B	3879
43	skokubezpieczenia24.moto2A	7757
44	skokubezpieczenia24.moto2B	7757
45	skokubezpieczenia24.moto3A	7757
46	skokubezpieczenia24.moto3B	7757
47	youcandrive.moto1A	1994
48	youcandrive.moto1B	1994
49	youcandrive.moto2A	1994
50	youcandrive.moto2B	1994
51	youcandrive.moto3A	1994
52	youcandrive.moto3B	1994
	RAZEM	275735

W rezultacie wzbogacenia danych nastąpił podział zbiorów na pierwotne i wzbogacone, co skutkowało zwiększeniem łącznej liczby rekordów do 275735. Liczebność poszczególnych zbiorów z uwzględnieniem poszczególnych podziałów oraz pierwotnych źródeł danych zestawiono w tabeli 13. W celu zwiększenia przejrzystość dalszego zarządzania zbiorami danych przyjęto następujące reguły co do oznaczeń zbiorów:

- pojedyncza cyfra na przedostatnim miejscu oznacza podzbiór wyodrębniony w ramach podziału w pierwszym wariancie (produkt ubezpieczeniowy),

- pojedyncza litera na końcu nazwy oznacza odpowiednio: A – zbiór danych niewzbogaconych, B - zbiór danych wzbogaconych.

Podsumowując tę cześć eksperymentu należy zauważyć, że łącznie uzyskano 52 zbiory danych będące wynikiem podwójnego podziału. Spośród tych 52 zbiorów 32 dotyczyły ubezpieczeń motoryzacyjnych. Z kolei jeśli chodzi o skalę pokrycia rynku, to zebrano dane z około 80% firm ubezpieczeniowych w Polsce udostępniających kalkulatory on-line (ale, nie wszystkich kalkulatorów).

Podczas zbierania danych doszło do szeregu zdarzeń, które odnotowane zostały w dzienniku badań, a związanych z odnośnymi źródłami WWW. Wśród istotniejszych z tych zdarzeń wskazać można:

- usunięcie kalkulatora (2 przypadki w trakcie i kolejne 2 po zakończeniu badań),
- wykryte zmiany w algorytmach (4 przypadki),
- wprowadzenie czasowej promocji (1 przypadek),
- zmiany w formularzu (7 przypadków),
- zmiana w bazie danych (1 przypadek).

Wszystkie powyższe zdarzenia zostały zaklasyfikowane jako problem rozciągłości zbierania danych ze źródeł w czasie. Uogólniając, na problem ten składa się nie tylko zmienność źródła – pojawiają się lub usuwane są zmienne decyzyjne, zmieniają zbiory wartości – ale również drugi czynnik – niektóre ubezpieczenia mogą być zależne od czasu. Ten ostatni czynnik wymusił dodatkowe zaimplementowanie odpowiednich mechanizmów w języku opisu źródła.

Kolejny zdiagnozowany problem, który pojawił się na tym etapie, to problem odpowiedniej interpretacji pustych danych. Z jednej strony mamy tutaj dane wadliwie pobrane[256], z drugiej puste zmienne warunkowe[257].

7.2 Założenia procedury ewaluacji

Materiał badawczy stanowiący wejście do procesu generowania modeli stanowią zbiory, których proces wyodrębniania opisano w podrozdziale 7.1. Zbiory te poddane

[256] O błędach w ekstrakcji mowa jest w dalszej części.
[257] Sytuacja, w której jedna zmienna niezależna nie ma przypisanej wartości, ponieważ jej występowanie jest zależne od wartości innej zmiennej niezależnej.

zostały modelowaniu za pomocą oprogramowania analitycznego przedstawionego w podrozdziale 6.5. W rezultacie dla każdego z 52 zbiorów danych wymienionych w podrozdziale 7.1 utworzono 23 modele objaśniające wielkość składki. Zestawienie wszystkich testowanych narzędzi analitycznych umożliwiających konstrukcję modeli wraz ze stosowanymi dalej oznaczeniami przedstawia tabela 14.

Tabela 14. Informacja o narzędziach (metodach) analitycznych wykorzystanych do konstrukcji modeli
Źródło: opracowanie własne

Oznaczenie narzędzia	Metoda	Komentarz
Neural	Sieć neuronowa	
Ensmbl	Złożenie	Model złożony z rezultatów pozostałych modeli nie opartych na głównych składowych
AutoNeural	AutoNeural	Automatycznie hodowana sieć neuronowa.
Tree	Drzewo decyzyjne	
DmineReg	Regresja DMINE	
PLS	Cząstkowe najmniejsze kwadraty	
Reg	Regresja	
DMNeural	DMNeural	
LARS	LARS	
MBR2	Metody pamięciowe	Model wykorzystujący rozkład na główne składowe
Boost	Boosting gradientowy	
Tree2	Drzewo decyzyjne	Model wykorzystujący rozkład na główne składowe
Boost2	Boosting gradientowy	Model wykorzystujący rozkład na główne składowe
Ensmbl2	Złożenie	Model złożony z rezultatów pozostałych modeli opartych na głównych składowych
DmineReg2	Regresja DMINE	Model wykorzystujący rozkład na główne składowe
AutoNeural2	AutoNeural	Model wykorzystujący rozkład na główne składowe
Neural2	Sieć neuronowa	Model wykorzystujący rozkład na główne składowe
PLS2	Cząstkowe najmniejsze kwadraty	Model wykorzystujący rozkład na główne składowe
Reg2	Regresja	Model wykorzystujący rozkład na główne skła-

		dowe
DMNeural2	DMNeural	Model wykorzystujący rozkład na główne składowe
LARS2	LARS	Model wykorzystujący rozkład na główne składowe
MBR	Metody pamięciowe	
PG	Programowanie genetyczne	Model generowany w ramach odrębnego oprogramowania

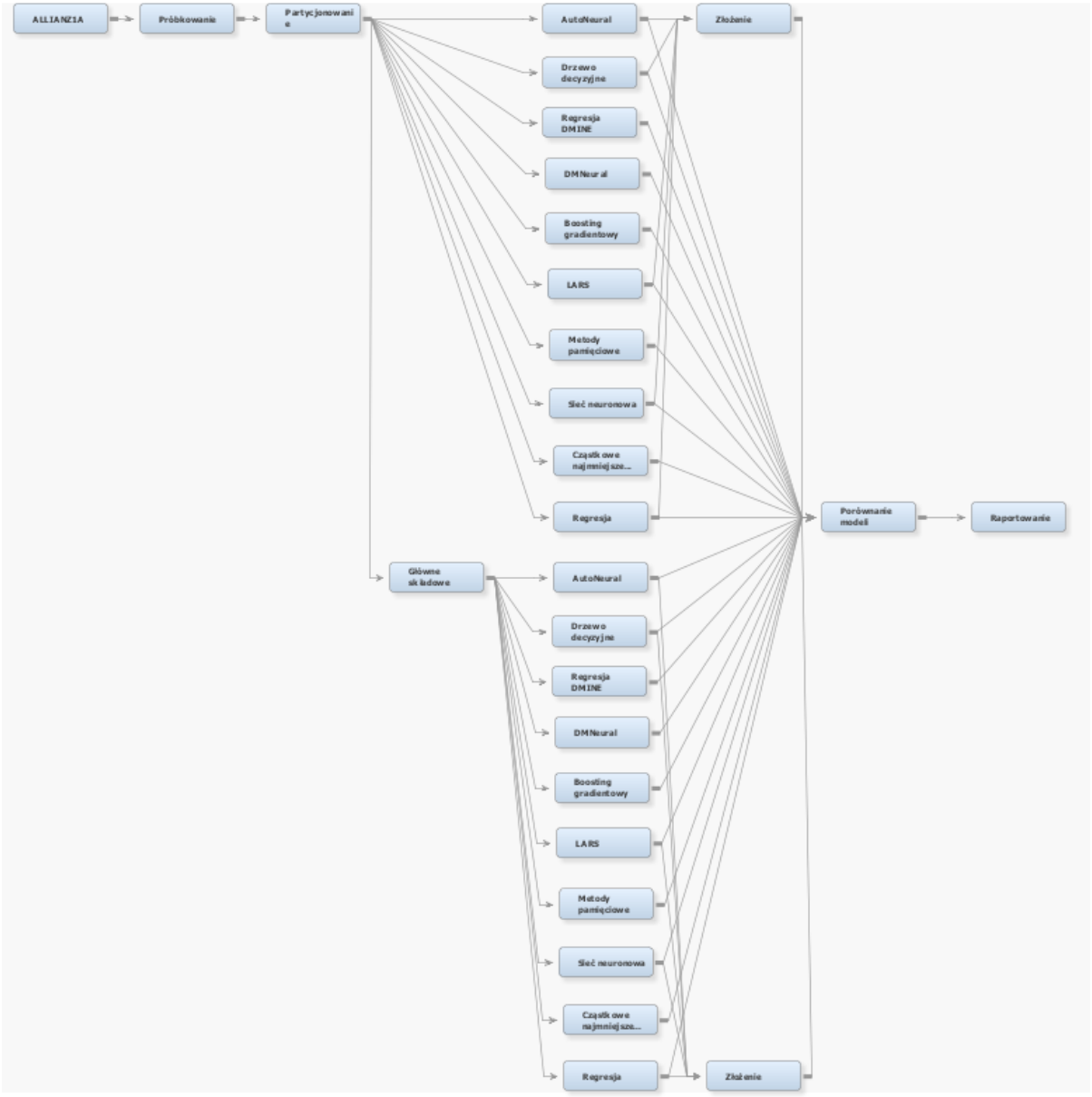
Część modeli ma charakter wtórny, objawiający się na dwa sposoby. Modele z nazwą zakończoną cyfrą „2" operują na przetworzonych zbiorach danych za pomocą procedury rozkładu na główne składowe. Z kolei w przypadku modeli, których nazwa zaczyna się od „Ensmbl", mamy do czynienia z modelem agregującym odpowiedzi pozostałych modeli.

Na rysunku 18 zademonstrowano schemat powiązań i przepływu danych pomiędzy poszczególnymi fragmentami procesu generowania modeli. Widać tam także powiązania wspomnianych modeli „Ensmbl" oraz „Ensmbl2" z pozostałymi modelami.

Dla każdego narzędzia analitycznego ustalono standardowe parametry generowania modeli oraz określono warunek zakończenia działania. Parametry te oraz warunki były jednakowe dla wszystkich badanych zbiorów danych. W przypadku narzędzi o charakterze procesu optymalizacji, miarą dobroci modelu była minimalizacja średniego błędu kwadratowego (MSE) walidacyjnego podzbioru danych.

7.3 Metoda oceny

Wprowadzono podział na jakościową i ilościową ocenę rezultatów eksperymentu. Podział ten traktujemy bardziej w kategorii wzajemnego uzupełnienia niż jako klasyczne przeciwstawienie dwóch metodologicznych punktów widzenia. W części poświęconej ocenie jakościowej staramy się podać informacje charakteryzujące jakość poszczególnych etapów eksperymentu. Wbrew klasycznemu rozumieniu pojęcia „oceny jakościowej", tam gdzie to możliwe owa jakość przedstawiana jest za pomocą parametrów liczbowych. Ze względu na przedmiot badań, taka parametryzacja jest po pierwsze możliwa, a po drugie niesie ze sobą więcej informacji. W rezultacie jako udokumentowanie poszczególnych prac podajemy:



Rysunek 18. Schemat powiązań i przepływu danych zastosowany do generowania modeli w systemie SAS
Źródło: SAS EM

- liczbę surowych rekordów i udział wadliwych rekordów – dokumentującą proces ekstrakcji danych,
- zestawienie porównania dwóch modeli z różnych, ale powiązanych źródeł – ze względu na dodatkową wiedzę o ich pochodzeniu, ich podobieństwo wskazuje na nielosowy i nieprzypadkowy charakter zebranych danych,
- listę i opis dodatkowych testów wzbogacających procedurę, których wyniki dokumentują przyjęte standardy pracy przy prowadzonym eksperymencie,
- podział zmiennych na typy w poszczególnych zbiorach danych oraz modelach – pozwalający ocenić poziom skomplikowania przedmiotu badań.

W dalszej części koncentrujemy się na końcowych wynikach eksperymentu, tj. na uzyskanych modelach wyceny składki. Modele te są charakteryzowane za pomocą typowych miar ilościowych – stąd ich przyporządkowanie do etapu ewaluacji ilościowej. W ramach tej ewaluacji przedstawiono:

- charakterystykę wielkości składek,
- zestawienie najlepszych modeli uzyskanych za pomocą oprogramowania SAS,
- zestawienie modeli wygenerowanych za pomocą oprogramowania do programowania genetycznego.

Tabela 15. Liczba surowych rekordów danych zebranych w procesie ekstrakcji z wyszczególnieniem adresów źródeł oraz podziałem na typy ubezpieczeń
Źródło: opracowanie własne

Źródło	Adres źródła	Typ	Liczba surowych rekordów
1	aviva.pl	turystyczne	2129
2	ehome.benefia24.pl	mieszkanie	2069
3	kuke.com.pl	handlowe	2723
4	signal-iduna.pl	życie	1202
5	skokubezpieczenia24.pl	mieszkanie	1842
6	skokubezpieczenia24.pl	życie	1091
7	tutum.pl	dla rowerzystów	365
8	tutum.pl	nnw	1728
9	uniqa24.pl	mieszkanie	2492
10	youcandrive.pl	turystyczne	2017
11	youcandrive.pl	mieszkanie	2136
12	allianzdirect.pl	moto	3061
13	axadirect.pl	moto	2382
14	emoto.benefia24.pl	moto	2393
15	libertydirect.pl (lu.pl)	moto	2607
16	link4.pl	moto	2572
17	mtusa.pl	moto	2505
18	skokubezpieczenia24.pl	moto	2800
19	youcandrive.pl	moto	1994
		RAZEM	40108

Obydwa zestawienia zostały sporządzone w oparciu o miarę średniego błędu kwadratowego (MSE). Stąd wartości tych błędów w odnośnych tabelach. Dodatkowo w przypadku drugiego zestawienia podano jako pomocniczą miarę współczynnik de-

terminacji (R^2). Na koniec dokonujemy agregacji powyższej informacji zestawiając ze sobą rankingi modeli utworzone według szeregu kryteriów.

7.4 Ewaluacja jakościowa

Zestawienie wszystkich zbiorów danych wraz z końcową ich wielkością znajduje się w podrozdziale 7.1. Aktualnie podajemy spis wykorzystanych źródeł oraz podział według typu ubezpieczenia przypadającej liczby rekordów surowych danych na konkretne źródło danych. Liczby surowych rekordów stanowią punkt wyjścia dla utworzenia poszczególnych zbiorów danych według podziałów opisanych w rozdziale 7.1.

Jak wynika z tabeli 15, eksperymentu dokonano na szerokim zakresie różnorodnych produktów ubezpieczeniowych. Przy czym największy udział miały ubezpieczenia motoryzacyjne. Taki dobór jest nieprzypadkowy, ponieważ ubezpieczenia te stanowią z jednej strony istotny segment całego rynku, z drugiej zaś są najpowszechniej oferowane w kanale sprzedaży internetowej.

Jeśli chodzi o proces ekstrakcji, to dla wszystkich źródeł WWW średnio 93,44% rekordów było prawidłowych. Dokładny rozkład efektywności ekstrakcji w rozbiciu na poszczególne źródła zebrano w tabeli 16. Rekordy wadliwe to takie, dla których zwrócone wartości były puste lub miały charakter nieliczbowy. Ponieważ ekstrakcja jest procesem automatycznym, więc nie istniał mechanizm pozwalający odróżnić błąd ekstrakcji od luk w modelu pierwotnym (niemożności wyceny składki po stronie ubezpieczyciela)[258].

W celu zapobieżenia i poradzenia sobie z niektórymi sytuacjami wyjątkowymi, o których mowa jest w podrozdziale 7.1, dodatkowo w stosunku do zaplanowanych prac rozwojowych opisanych w rozdziale 6.5, zaimplementowano specjalne testy jednostkowe. Testy te miały dostarczyć następujące funkcjonalności:

- test dostępności wszystkich źródeł,
- test kompletności cyklu dla wybranego źródła,
- test kompletności cyklu dla wszystkich źródeł,
- test niezmienności algorytmu dla wybranego źródła,

[258] W sytuacjach ewidentnych takie przypadki były wyłapywane manualnie. Oczywiście, można sobie wyobrazić automatyczne rozwiązanie takich sytuacji, pozostało to jednak poza obszarem eksperymentu.

- test niezmienności algorytmu dla wszystkich źródeł.

Tabela 16. Udział prawidłowych i nieprawidłowych rekordów otrzymanych w procesie ekstrakcji z wyszczególnieniem źródeł danych
Źródło: opracowanie własne

Źródło	% prawidłowych	% wadliwych
aviva1A/B	98,00%	2,00%
ehome.benefia241A	95,28%	4,72%
kuke1A/B	99,30%	0,70%
signal-iduna1A/B	94,84%	5,16%
skokubezpieczenia24.home1A	99,27%	0,73%
skokubezpieczenia24.home2A	96,90%	3,10%
skokubezpieczenia24.health1A	70,33%	29,67%
tutum.bike1A	98,93%	1,07%
tutum.nnw1A	76,86%	23,14%
uniqa241A	96,07%	3,93%
uniqa242A	96,07%	3,93%
youcandrive.travel1A/B	95,39%	4,61%
youcandrive.travel2A/B	80,06%	19,94%
allianz1A/B	100,00%	0,00%
allianz2A/B	100,00%	0,00%
axadirect1A/B	86,39%	13,61%
axadirect2A/B	77,57%	22,43%
emoto.benefia241A/B	98,83%	1,17%
libertydirect1A/B	92,77%	7,23%
libertydirect2A/B	98,72%	1,28%
link41A/B	92,00%	8,00%
link42A/B	92,03%	7,97%
mtusa1A/B	99,70%	0,30%
skokubezpieczenia24.moto1A/B	98,94%	1,06%
skokubezpieczenia24.moto2A/B	91,62%	8,38%
skokubezpieczenia24.moto3A/B	91,59%	8,41%
youcandrive.moto1A/B	99,40%	0,60%
youcandrive.moto2A/B	96,08%	3,92%
youcandrive.moto3A/B	96,68%	3,32%
ŚREDNIA	93,44%	6,56%

W konsekwencji wprowadzenia powyższych funkcjonalności testy powinny gwarantować w związku z procesem ekstrakcji wykrycie: niedostępności źródła, zmian w kształcie formularzy oraz niestabilności modelu.

Udało się także przeprowadzić eksperyment porównawczy bazujący na danych pomiędzy źródłami mtusa.pl oraz skokubezpieczenia24.pl (ubezpieczenie motoryza-

cyjne). Ostatni z ubezpieczycieli informuje na stronie internetowej, że oferta produktowa przygotowana została we współpracy z pierwszym z ubezpieczycieli. MTU S.A. jest towarzystwem ubezpieczeniowym o znacznie dłuższej tradycji na rynku, a co za tym idzie, także o zdecydowanie większym doświadczeniu w oferowaniu ubezpieczeniowych produktów motoryzacyjnych[259]. Postawiono zatem hipotezę badawczą, że oferty winny być w istotny sposób skorelowane lub wręcz tożsame. Ewentualnie mogłyby się one różnić o marżę drugiego z ubezpieczycieli. Wykazanie prawdziwości hipotezy w istotny sposób uwiarygodniałoby zarówno poprawność zebranych danych, jak też stanowiłoby istotną przesłankę do pozytywnej oceny wygenerowanych modeli.

Po przeprowadzeniu analizy tych dwóch modeli okazało się, że są one zgodne przy uwzględnieniu prostych reguł przekształceń. Drugi z modeli różni się w stosunku do pierwszego o zmienną marżę, która jest wyłącznie zależna od regionu geograficznego. Przeprowadzona analizy dała zatem wynik pozytywny. Przykładowe wykryte różnice pomiędzy modelami zestawiono w tabeli 17. Brak zaokrągleń w procentowej różnicy składek wynika z faktu, że pierwsze źródło podaje wartości składek już zaokrąglone do pełnych złotych.

Tabela 17. Wykryte różnice w poziomach składki pomiędzy modelami opartymi na źródłach mtusa.pl oraz skokubezpieczenia24.pl
Źródło: opracowanie własne

Miejscowość	Kod pocztowy regionu	Różnica składki
Bełchatów	97-400	4,96%
Warszawa	03-112	2,93%
Dębica	39-202	7,13%
Gniezno	62-200	5,02%
Kołobrzeg	78-100	4,99%
Lubin	59-300	1,84%
Łódź	94-020	5,04%
Biała Podlaska	21-500	5,01%
Bydgoszcz	85-455	-0,24%
Chorzów	41-519	7,13%
Elbląg	82-300	3,98%
Gliwice	44-100	4,95%
Jastrzębie-Zdrój	44-335	1,88%
Kalisz	62-800	4,95%
Konin	62-500	5,01%

[259] Sytuacje tego typu omówiono od strony teoretycznej w rozdziale 3.3.

Krosno	38-400	10,27%
Lubin	59-304	1,84%
Mysłowice	41-401	7,09%
Opole	45-076	4,94%
Piotrków Trybunalski	97-304	5,00%
Przemyśl	37-707	7,15%
Rybnik	44-200	-0,24%
Siemianowice Śląskie	41-100	5,02%
Sosnowiec	41-219	7,06%
Szczecin	70-840	2,57%
Tarnobrzeg	39-400	2,58%
Tychy	43-102	3,49%
Zabrze	41-806	2,58%
Żory	44-240	3,48%
Nysa	48-300	2,59%
Ostrów Wielkopolski	63-400	2,20%
Piaseczno	05-500	3,30%
Poznań	60-660	3,30%
Poznań	61-530	3,30%
Pruszków	05-809	2,57%
Radomsko	97-500	2,58%
Skarżysko-Kamienna	26-110	3,12%
Stargard Szczeciński	73-100	2,57%
Świdnica	58-100	2,58%
Tczew	83-101	2,59%
Warszawa	02-783	2,93%
Wręczyca Mała	42-130	6,35%
Wrocław	50-065	9,72%
Zduńska Wola	98-220	-0,24%
Żyrardów	96-300	1,55%
Lublin	20-105	3,74%
Kraków	30-109	-0,24%
Wejherowo	84-200	4,37%

Stworzone zbiory danych przeznaczone do przetworzenia na modele składały się ze zróżnicowanych typów danych. Podstawowe dwa typy to dane ciągłe (interwałowe) oraz dane nominalne, inaczej dyskretne (o charakterze liczbowym lub tekstowym). W przypadku występowania danych binarnych – logicznych, zamieniane były one na dane o typie nominalnym liczbowym. Zestawienie liczebności typów danych zawiera tabela 18.

Na koniec podrozdziału pokazujemy fragment otrzymanego kodu modelu wyhodowanego za pomocą metody programowania genetycznego (rysunek 19). Programy mają długość rzędu 1024 instrukcji.

Tabela 18. Zestawienie liczby rodzajów zmiennych niezależnych w podziale na źródła danych
Źródło: opracowanie własne

Zbiór danych	Zm. czasowe	Interwałowe	Nominalne	
			Binarne	Wielowartościowe
aviva1A	2	13	3	1
aviva1B	2	17	3	1
ehome.benefia241A	0	21	4	1
kuke1A.com	0	3	0	1
kuke1B.com	0	14	0	1
signal-iduna1A	1	10	2	16
signal-iduna1B	1	23	2	16
skokubezpieczenia24.home1A	2	39	2	2
skokubezpieczenia24.home2A	2	39	2	2
skokubezpieczenia24.health1A	2	16	3	2
tutum.bike1A	3	8	2	2
tutum.nnw1A	3	9	2	2
uniqa241A	0	25	4	2
uniqa242A	0	26	4	2
youcandrive.home1A	1	14	12	4
youcandrive.home1B	1	17	12	5
youcandrive.travel1A	3	8	16	4
youcandrive.travel1B	3	11	16	4
youcandrive.travel2A	3	8	16	4
youcandrive.travel2B	3	11	16	4
allianz1A	2	8	4	8
allianz1B	2	17	4	12
allianz2A	2	8	4	8
allianz2B	2	17	4	12
axadirect1A	2	38	7	39
axadirect1B	2	44	7	42
axadirect2A	2	39	7	39
axadirect2B	2	45	7	42
emoto.benefia241A	0	13	3	7
emoto.benefia241B	0	21	3	10
libertydirect1A	1	35	5	6
libertydirect1B	1	43	5	9
libertydirect2A	1	34	5	6
libertydirect2B	1	42	5	9
link41A	2	36	7	25
link41B	2	44	7	8
link42A	2	37	7	25
link42B	2	45	7	8
mtusa1A	1	9	3	7
mtusa1B	1	16	3	9
skokubezpieczenia24.moto1A	1	42	4	11

skokubezpieczenia24.moto1B	1	49	4	13
skokubezpieczenia24.moto2A	1	43	4	11
skokubezpieczenia24.moto2B	1	50	4	13
skokubezpieczenia24.moto3A	1	43	4	11
skokubezpieczenia24.moto3B	1	50	4	13
youcandrive.moto1A	2	17	3	12
youcandrive.moto1B	2	25	3	13
youcandrive.moto2A	2	17	3	12
youcandrive.moto2B	2	25	3	13
youcandrive.moto3A	2	17	3	12
youcandrive.moto3B	2	25	3	13
ŚREDNIA	1,534	25,5	5,13	10,65
UDZIAŁ	3,59%	59,54%	11,99%	24,88%

Jak widać ich konstrukcja ma dość chaotyczny charakter, co jednak jest immanentną cechą tego typu modeli wynikającą ze sposobu ich budowania.

```
f[2]+=f[0]; f[0]+=-0.1981618404388428f; f[0]/=use-type;
tmp=f[0]; f[0]=f[0]; f[0]=tmp; f[0]=Math.sqrt(f[0]);
f[0]+=0.5301053524017334f; f[0]=Math.cos(f[0]);
f[0]/=f[0]; f[0]+=spec_vehicle;
f[0]*=f[2]; f[2]+=f[0];
f[0]+=m-car-engine-capacity-list; f[0]+=m-car-engine-capacity-list;
f[1]+=f[0]; f[0]-=m-car-fuel-type-list;
if (!cflag) f[0] = f[1]; f[2]+=f[0];
f[0]=Math.cos(f[0]); f[1]-=f[0];
f[0]+=average-mileage; f[0]-=f[1];
f[0]*=use-type; f[0]-=spec_holderpostcodecity;
f[0]=Math.abs(f[0]); f[0]*=f[0];
f[0]/=m-car-fuel-type-list; tmp=f[0]; f[0]=f[0]; f[0]=tmp;
f[0]*=m-car-engine-capacity-list; f[0]=Math.cos(f[0]);
f[2]+=f[0]; f[0]=Math.abs(f[0]); f[0]*=f[0];
f[0]*=date-of-birth-year; f[1]+=f[0];
```

Rysunek 19. Fragment kodu modelu otrzymanego za pomocą metody programowania genetycznego
Źródło: opracowanie własne

7.5 Ewaluacja ilościowa

W ramach eksperymentu, jak już wspomniano, dla każdego z 52 zbiorów danych utworzono 23 modele objaśniające wielkość składki. Podstawowe parametry opisujące dane wejściowe do generowanych modeli zostały zebrane w tabeli 19. Ponieważ parametry charakteryzują samą składkę, która nie była różnicowana w podziale danych w drugim wariancie (wzbogacenie o dane zewnętrzne) zatem podzbiory A i B zestawiono łącznie. Wszystkie podane wartości w tabeli denominowane są w złotych polskich.

Tabela 19. Ogólna charakterystyka danych (wielkości składek) zebranych w trakcie eksperymentu w rozbiciu na poszczególne źródła
Źródło: opracowanie własne

Źródło danych	Minimum	Maksimum	Rozpiętość	Średnia	Odchylenie standard.
aviva1A/B	18	302	284	89,39151	51,73615
ehome.benefia241A	100	1399	1299	516,8171	206,4825
kuke1A/B	315	252300	251985	75220,12	82895,42
signal-iduna1A/B	79	5644,8	5565,8	591,8688	728,5648
skokubezpieczenia24.home1A	61	1006	945	346,5227	187,4
skokubezpieczenia24.home2A	65	1066	1001	417,1743	208,1082
skokubezpieczenia24.health1A	54	392	338	149,1458	66,10796
tutum.bike1A	4,6	180	175,4	38,00255	33,23091
tutum.nnw1A	28	2078	2050	295,716	324,0922
uniqa241A	109	320	211	134,0008	44,5205
uniqa242A	14	833	819	198,9243	156,5804
youcandrive.home1A/B	10	1077	1067	412,1226	201,9232
youcandrive.travel1A/B	12	11051	11039	944,1397	1278,431
youcandrive.travel2A/B	5	3402	3397	234,116	337,2447
allianz1A/B	50	12327	12277	1544,951	1554,928
allianz2A/B	50	10768	10718	2122,666	1731,241
axadirect1A/B	429	1619	1190	833,9927	169,907
axadirect2A/B	479	4143	3664	1536,356	437,6497
emoto.benefia241A/B	460	5787	5327	1533,569	1051,358
libertydirect1A/B	695,64	11927,08	11231,44	2104,974	1547,548
libertydirect2A/B	601,6	3428,33	2826,73	1103,235	489,3529
link41A/B	19,92	5113,21	5093,29	2331,275	539,8021
link42A/B	2	11340,47	11338,47	3700,05	1288,632
mtusa1A/B	718	3055	2337	1458,606	389,1469
skokubezpieczenia24.moto1A/B	547,71	5477,08	4929,37	2123,657	1558,239
skokubezpieczenia24.moto2A/B	1001,78	11413,6	10411,82	3948,972	2190,562
skokubezpieczenia24.moto3A/B	297,72	9975,68	9677,96	2209,335	1906,452
youcandrive.moto1A/B	643	2933	2290	1153,216	376,18
youcandrive.moto2A/B	350	5069	4719	1122,162	769,8174
youcandrive.moto3A/B	541	8027	7486	2390,535	1482,011

Obecnie zaprezentujemy zestawienie najlepszych modeli uzyskanych w ramach oprogramowania analitycznego SAS. W tabeli 20 podano średni błąd kwadratowy odpowiednio dla podzbioru danych treningowych oraz walidacyjnych dla każdego z analizowanych zbiorów. Podzbiór treningowy, to dane wyodrębnione losowo z całkowitego zbioru danych w celu zbudowania modelu. Z kolei podzbiór danych walidacyjnych jest tworzony na analogicznej zasadzie w celu oceny jakości modelu. Dla oceny modelu właściwszy jest podzbiór walidacyjny, gdyż zawiera nowe dane w stosunku

do danych, w oparciu o które model był budowany. Ponieważ modele generowane są w oparciu o ten pierwszy podzbiór danych, należy oczekiwać, że optymalizowana miara będzie nieco niższa właśnie dla tego podzbioru. Uzyskane wyniki są zatem zgodne z oczekiwaniami. Dodatkowo zbliżenie wielkości błędu uzyskanych dla obu podzbiorów może być (pośrednio) miarą jakości modelu. Mała zmiana w otrzymanym błędzie w próbie walidacyjnej w stosunku do próby treningowej świadczy o tym, że model daje zbliżone wyniki dla potencjalnie dowolnych danych wejściowych.

Tabela 20. Miary obrazujące jakość otrzymanych modeli stworzonych za pomocą systemu SAS. Zestawienie nie obejmuje programowania genetycznego
Źródło: opracowanie własne

Zbiór danych	Najlepszy model	MSE trening	MSE walidacja
allianz1A	Neural	715929,04	927281,52
allianz1B	Tree	781993,85	1031951,62
allianz2A	DmineReg	253638,19	203726,55
allianz2B	PLS	233479,44	183238,6
aviva1A	AutoNeural2	8,216	13,876
aviva1B	Tree	77,53	91,92
axadirect1A	Ensmbl	1059,65	1076,41
axadirect1B	Ensmbl	958,98	1046,09
axadirect2A	AutoNeural	8452,55	9576,6
axadirect2B	AutoNeural	7291,31	7854,9
ehome.benefia241A	Neural2	397,44	360,3
emoto.benefia241A	Tree	5844,64	9070,1
emoto.benefia241B	Tree	5844,64	9070,1
kuke1A	Tree	339822193,7	428866297,5
kuke1B	Tree	302319,04	297433,75
libertydirect1A	Tree	68303,87	64766,32
libertydirect1B	Tree	68105,08	65430,22
libertydirect2A	Neural	2182,52	2643,51
libertydirect2B	DmineReg	2416,77	2724,92
link41A	Neural	11656,16	9619,67
link41B	Ensmbl	66725,9	60021,69
link42A	Neural	45722,63	47853,89
link42B	Neural	131215,64	127991,71
mtusa1A	Neural	9583,68	10099,51
mtusa1B	DmineReg	12278,61	10980,74
signal-iduna1A	Tree	16029,1	35439,85
signal-iduna1B	Tree	16029,1	35439,85
skokubezpieczenia24.health1A	Tree	66,32	66,65
skokubezpieczenia24.home1A	Tree	887,99	914,65
skokubezpieczenia24.home2A	LARS	87,56	122,34

skokubezpieczenia24.moto1A	Neural	21015,03	25713,57
skokubezpieczenia24.moto1B	Neural	21188,87	27005,44
skokubezpieczenia24.moto2A	AutoNeural	84650,13	113127,6
skokubezpieczenia24.moto2B	Neural	61372,09	92526,61
skokubezpieczenia24.moto3A	Tree	95068,83	190134,43
skokubezpieczenia24.moto3B	Tree	94336,91	185935,16
tutum.bike1A	Tree	1,82	2,41
tutum.nnw1A	AutoNeural	1907,95	2810,66
uniqa241A	Tree	2,29	1,98
uniqa242A	MBR	2494,38	3066,18
youcandrive.home1A	LARS	65,64	57,44
youcandrive.home1B	LARS	66,06	57,64
youcandrive.travel1A	Tree	114314,16	120255,59
youcandrive.travel1B	Tree	117733,46	127496,84
youcandrive.travel2A	Tree	6821,32	7575,29
youcandrive.travel2B	Tree	9455,55	9228,89
youcandrive.moto1A	Neural	65844,68	72560,32
youcandrive.moto1B	Tree	65327,00	72042,64
youcandrive.moto2A	DmineReg	11533,53	16371,72
youcandrive.moto2B	PLS	11048,42	15886,61
youcandrive.moto3A	DmineReg	15369,85	20208,04
youcandrive.moto3B	Neural	68387,92	75103,56

W tabeli 20 wyselekcjonowano dokładnie jeden model dla każdego badanego zbioru danych. Wbrew naszym oczekiwaniom, spośród powyżej zaprezentowanych narzędzi analitycznych najlepszym okazały się drzewa decyzyjne (oznaczenie Tree). Dopiero na drugim miejscu znalazły się klasyczne sieci neuronowe.

Kolejne dwie tabele (numer 21 i 22) pozwalają na porównanie średnich błędów kwadratowych (MSE) i współczynnika determinacji (R^2) dla modeli otrzymanych za pomocą metody programowania genetycznego na poszczególnych zbiorach danych[260].

Przewagą programowania genetycznego jest otrzymanie analitycznej postaci modelu (przykład dla jednego z takich modeli zamieszczono w podrozdziale 7.4). Podana liczba wygenerowanych programów powstała i została przetestowana w procesie ewolucji w celu poszukiwania najlepszego, podanego wyniku.

[260] Miary policzono osobno dla podzbioru danych treningowych oraz walidacyjnych.

Tabela 21. Miary obrazujące jakość otrzymanych modeli stworzonych za pomocą metody programowania genetycznego (pojedyncze programy) wraz z liczbą wszystkich przetestowanych programów
Źródło: opracowanie własne

Zbiór danych	MSE walidacja	MSE trening	R^2 walidacja	R^2 trening	Liczba programów
allianz1A	447286,75	359857,63	82,12%	85,67%	20350759
allianz1B	410451,50	471969,19	84,85%	81,74%	21026066
allianz2A	520377,94	379336,84	83,26%	86,73%	25942737
allianz2B	507949,31	484823,94	82,04%	84,24%	22905427
aviva1A	106,50	113,12	95,91%	95,97%	54752759
aviva1B	60,19	58,47	97,81%	97,70%	45647374
axadirect1A	11946,31	12682,11	58,13%	52,29%	20081554
axadirect1B	8249,61	7411,37	71,11%	73,75%	23783640
axadirect2A	65453,20	64645,83	67,16%	62,78%	31820204
axadirect2B	56629,39	59140,88	69,70%	69,05%	22844466
ehome.benefia241A	1994,37	1984,35	95,34%	95,31%	20555394
emoto.benefia241A	48799,38	44211,49	96,12%	95,64%	24368268
emoto.benefia241B	39112,48	43508,46	96,33%	96,63%	20516398
kuke1A	4343913,00	4325496,50	93,65%	93,74%	24004061
kuke1B	3153514,25	3003570,25	95,25%	95,78%	23791071
libertydirect1A	179647,02	179143,56	92,99%	92,59%	21280548
libertydirect1B	161821,34	162212,00	93,49%	93,17%	30902374
libertydirect2A	16584,68	17090,26	93,92%	91,82%	32062929
libertydirect2B	10338,06	10218,07	95,38%	95,67%	30648467
link41A	84542,02	80730,09	70,77%	73,75%	22307562
link41B	90990,57	86652,16	68,15%	70,62%	24272251
link42A	339069,06	349959,88	79,66%	77,75%	21432407
link42B	256692,97	248821,47	84,46%	84,94%	22249395
mtusa1A	80854,20	88977,86	48,05%	46,73%	20767416
mtusa1B	38855,17	26508,76	76,27%	81,50%	21229426
signal-iduna1A	10679,03	10104,01	97,59%	98,33%	26233498
signal-iduna1B	8890,81	6266,44	98,42%	97,59%	21614243
skokubezpieczenia24.health1A	163,64	151,93	96,44%	96,37%	27763175
skokubezpieczenia24.home1A	1365,04	1423,01	96,33%	96,22%	22656195
skokubezpieczenia24.home2A	88,00	78,35	99,80%	99,81%	50805405
skokubezpieczenia24.moto1A	129182,31	148556,64	94,78%	94,02%	28197666
skokubezpieczenia24.moto1B	114541,61	109980,67	95,64%	95,22%	21299320
skokubezpieczenia24.moto2A	456648,38	466885,72	90,41%	90,56%	29238616
skokubezpieczenia24.moto2B	668268,38	603866,94	86,32%	87,11%	21094867
skokubezpieczenia24.moto3A	278538,59	242568,03	92,28%	93,28%	42939614
skokubezpieczenia24.moto3B	296301,44	283995,13	90,92%	92,39%	22986749
tutum.bike1A	6,30	5,54	99,46%	99,49%	22690077
tutum.nnw1A	8223,37	7859,10	92,28%	92,59%	26557181
uniqa241A	6,69	6,75	99,82%	99,85%	22531604
uniqa242A	23889,29	23841,18	70,31%	70,60%	21316644

youcandrive.home1A	310,39	321,09	99,25%	99,28%	23094885
youcandrive.home1B	192,80	205,36	99,53%	99,49%	25456673
youcandrive.travel1A	128470,97	136868,09	92,27%	91,78%	21329856
youcandrive.travel1B	158288,27	117712,04	91,15%	91,17%	24557373
youcandrive.travel2A	9553,40	8563,06	91,58%	93,34%	22801585
youcandrive.travel2B	10573,76	11850,19	88,98%	90,32%	23210090
youcandrive.moto1A	15836,05	10997,86	95,32%	95,89%	33837199
youcandrive.moto1B	15318,37	10480,18	95,45%	95,99%	27689116
youcandrive.moto2A	152760,76	150694,49	88,24%	87,56%	32162478
youcandrive.moto2B	152275,65	150209,38	88,83%	90,23%	29532814
youcandrive.moto3A	156597,08	154530,81	88,64%	89,36%	53674518
youcandrive.moto3B	18379,29	13541,10	94,62%	95,17%	34480124
ŚREDNIO	264504,73	254826,88	88,13%	88,35%	27093583,86
RAZEM					1408866361

Przyjęta metoda programowania genetycznego umożliwia, poza wyborem poszczególnych najlepszych kandydatów – programów, także wybór rozwiązań zbiorowych. Polega to na doborze wyselekcjonowanych programów z aktualnej populacji i stworzeniu za ich pomocą modelu łączącego wyniki poszczególnych z nich[261]. Zakłada się, że takie złożone modele mogą mieć lepszą zdolność predykcyjną od pojedynczych elementów z populacji. W tabeli 22 podano zatem wyniki dla modeli zbiorowych.

Tabela 22. Miary obrazujące jakość otrzymanych modeli stworzonych za pomocą programowania genetycznego (najlepsze drużyny) wraz z przybliżonym czasem ich tworzenia
Źródło: opracowanie własne

Zbiór danych	MSE walidacja	MSE trening	R^2 walidacja	R^2 trening	Czas hodowania
allianz1A	384480,97	298020,97	84,63%	88,14%	00:29:08
allianz1B	293926,47	351406,78	89,15%	86,41%	00:25:31
allianz2A	409835,91	310806,19	86,81%	89,12%	02:08:03
allianz2B	380129,91	396000,38	86,56%	87,12%	00:25:20
aviva1A	73,34	76,22	97,18%	97,28%	01:48:35
aviva1B	52,62	51,78	98,08%	97,97%	02:22:13
axadirect1A	11491,59	11344,10	59,72%	57,33%	00:34:24
axadirect1B	6957,99	6414,99	75,64%	77,27%	00:40:51
axadirect2A	48039,48	49987,57	75,90%	71,22%	00:53:08
axadirect2B	45650,97	47556,95	75,57%	75,11%	00:31:36
ehome.benefia241A	1738,28	1701,20	95,94%	95,98%	01:28:47
emoto.benefia241A	46886,29	42409,19	96,28%	95,82%	00:44:56
emoto.benefia241B	34377,68	36730,44	96,77%	97,16%	00:34:18

[261] Jest to rozwiązanie zbliżone do modeli oznaczonych jako „Ensmbl" generowanych za pomocą systemu SAS.

kuke1A	65744540,00	67451392,00	3,88%	2,30%	01:23:15
kuke1B	47984652,00	56897856,00	27,67%	20,08%	00:39:27
libertydirect1A	131061,68	134376,78	94,89%	94,44%	01:17:26
libertydirect1B	125605,27	126556,29	94,95%	94,67%	01:23:07
libertydirect2A	14855,54	14955,74	94,55%	92,84%	01:56:26
libertydirect2B	9862,56	10212,60	95,59%	95,67%	01:11:30
link41A	77924,51	75995,31	73,06%	75,29%	01:20:09
link41B	78178,05	73161,15	72,64%	75,20%	01:06:02
link42A	298311,44	306880,75	82,10%	80,49%	01:56:06
link42B	204258,17	208361,36	87,63%	87,39%	00:53:20
mtusa1A	75670,21	81132,41	51,38%	51,42%	00:57:58
mtusa1B	35836,38	23733,97	78,12%	83,44%	00:39:30
signal-iduna1A	138,27	127,89	96,99%	96,94%	00:39:39
signal-iduna1B	6913,19	4782,91	98,77%	98,16%	00:45:23
skokubezpieczenia24.health1A	8330,78	8352,90	98,12%	98,62%	01:01:08
skokubezpieczenia24.home1A	1006,86	1017,32	97,29%	97,30%	01:15:13
skokubezpieczenia24.home2A	79,78	74,05	99,82%	99,82%	01:48:24
skokubezpieczenia24.moto1A	112017,98	119979,16	95,47%	95,17%	00:57:25
skokubezpieczenia24.moto1B	92379,27	83056,18	96,48%	96,39%	00:54:22
skokubezpieczenia24.moto2A	423396,53	453812,78	91,11%	90,82%	01:50:47
skokubezpieczenia24.moto2B	619437,00	537369,81	87,32%	88,53%	00:53:38
skokubezpieczenia24.moto3A	206833,23	185442,59	94,27%	94,86%	01:36:58
skokubezpieczenia24.moto3B	239523,97	219141,59	92,66%	94,13%	00:39:08
tutum.bike1A	4,37	4,31	99,63%	99,61%	00:28:28
tutum.nnw1A	6718,12	5892,04	93,70%	94,44%	00:21:13
uniqa241A	3,92	4,02	99,90%	99,91%	02:03:14
uniqa242A	16137,65	15251,34	79,95%	81,19%	02:03:14
youcandrive.home1A	290,37	291,69	99,30%	99,28%	01:32:36
youcandrive.home1B	181,20	189,82	99,56%	99,53%	01:56:04
youcandrive.travel1A	110306,48	118299,69	93,36%	92,90%	01:49:27
youcandrive.travel1B	153426,50	120407,34	91,42%	90,97%	01:10:48
youcandrive.travel2A	5548,10	5184,84	95,11%	95,97%	02:09:27
youcandrive.travel2B	7799,32	8158,98	91,87%	93,34%	02:05:20
youcandrive.moto1A	17736,37	12317,60	93,42%	93,97%	01:08:54
youcandrive.moto1B	17156,57	11737,80	93,54%	94,07%	01:10:41
youcandrive.moto2A	171092,05	168777,83	86,47%	85,81%	02:01:52
youcandrive.moto2B	170548,73	168234,51	87,06%	88,43%	01:15:05
youcandrive.moto3A	175388,73	173074,51	86,86%	87,58%	02:01:12
youcandrive.moto3B	20584,80	15166,03	92,72%	93,27%	00:58:42
ŚREDNIO	2289218,44	2488544,12	86,66%	86,72%	01:14:03

Jak łatwo można zauważyć, porównując wyniki z pierwszej tabeli modeli generowanych za pomocą metody programowania genetycznego oraz tabeli z wynikami mo-

deli zbiorczych, rzeczywiście te drugie modele przeważnie osiągają nieco lepsze rezultaty. Zarówno populacja programów podstawowych, jak i tworzenie modeli zbiorczych następują w trakcie tego samego procesu, w związku z czym podane czasy hodowania dotyczą czasu potrzebnego do stworzenia i przetestowania liczby programów podanych w pierwszej tabeli oraz dodatkowo stworzenia i przetestowania modeli zbiorowych.

W przypadku przebadanych zbiorów danych zastosowana metoda okazała się wysoce skuteczna. Obecnie zajmiemy się wskazaniem najlepszego narzędzia analitycznego do odtworzenia modeli wyceny składki w oparciu o wcześniej podane oraz pozostałe[262] wyniki cząstkowe.

Tabela 23. Ranking wszystkich metod analitycznych wg liczby punktów otrzymanych za miejsca zdobyte wg dopasowania do poszczególnych zbiorów danych (22 punkty 1-sze miejsce; 0 punktów – ostatnie miejsce)
Źródło: opracowanie własne

L.p.	Narzędzie analityczne	Ocena punktowa
1	Tree	999
2	PG	963
3	Ensmbl	914
4	Neural	913
5	DmineReg	884
6	LARS	784
7	Reg	714
8	PLS	701
9	Boost	605
10	AutoNeural	590
11	DMNeural	576
12	Ensmbl2	542
13	Neural2	520
14	MBR2	503
15	Tree2	502
16	MBR	474
17	DmineReg2	472
18	AutoNeural2	349
19	Boost2	310
20	Reg2	263
21	LARS2	214
22	PLS2	200
23	DMNeural2	164

[262] Ze względu na ograniczenia objętościowe wcześniej podaliśmy tylko wyniki najlepszych modeli, pomijając pozostałe wyniki.

W tabeli 23 zestawiono ranking poszczególnych metod analitycznych według liczby zdobytych punktów. Ocena punktowa jest sumą punktów zdobytych przez daną metodę dla kolejnych zbiorów danych. Punkty przyznawane były za miejsce zajęte spośród wszystkich narzędzi analitycznych dla danego zbioru danych. Dana metoda w przypadku pojedynczego zbioru danych mogła zdobyć maksymalnie 22 punkty za pierwsze miejsce oraz minimalnie 0 punktów za miejsce ostatnie. Miejsca ustalane były za pomocą rosnącego średniego błędu kwadratowego uzyskanego na podzbiorach walidacyjnych.

Następne trzy rankingi powstały ze złożenia informacji zawartej w tabelach od 20 do 22. Porównują one odpowiednio liczbę zajętych pierwszych miejsc dla poszczególnych modeli i zbiorów danych. Uwzględniają one zarówno modele uzyskane za pomocą programowania genetycznego oraz wszystkie pozostałe.

Tabela 24. Ranking wszystkich metod analitycznych wg liczby zajęcia pierwszego miejsca dla poszczególnych zbiorów danych
Źródło: opracowanie własne

L.p.	Narzędzie analityczne	Liczba najlepszych modeli (łącznie)
1	PG	13
2	Tree	12
3	Neural	8
4	DmineReg	5
5	AutoNeural	4
6	Ensmbl	3
7	LARS	2
8	PLS	2
9	AutoNeural2	1
10	MBR	1
11	Neural2	1
12	Boost	0
13	Boost2	0
14	DmineReg2	0
15	DMNeural	0
16	DMNeural2	0
17	Ensmbl2	0
18	LARS2	0
19	MBR2	0
20	PLS2	0
21	Reg	0
22	Reg2	0
23	Tree2	0

Podstawowa różnica pomiędzy rankingami zawartymi w tabelach od 24 do 26 a poprzednim rankingiem polega na tym, że poprzedni ranking uwzględniał także wyniki słabiej radzących sobie metod analitycznych.

Tabela 24 przedstawia ranking stworzony w oparciu o wszystkie zbiory danych łącznie, podczas gdy tabele 25 i 26 na tej samej zasadzie dokonują zestawienia tylko dla podzbiorów, odpowiednio, A (dane niewzbogacane) oraz B (dane wzbogacone).

Tabela 25. Ranking wszystkich metod analitycznych wg liczby zajęcia pierwszego miejsca dla podzbioru danych A (dane niewzbogacone)
Źródło: opracowanie własne

L.p.	Narzędzie analityczne	Liczba najlepszych modeli (podzbiór A)
1	PG	7
2	Tree	7
3	Neural	5
4	AutoNeural	3
5	DmineReg	3
6	AutoNeural2	1
7	Ensmbl	1
8	LARS	1
9	MBR	1
10	Neural2	1
11	Boost	0
12	Boost2	0
13	DmineReg2	0
14	DMNeural	0
15	DMNeural2	0
16	Ensmbl2	0
17	LARS2	0
18	MBR2	0
19	PLS	0
20	PLS2	0
21	Reg	0
22	Reg2	0
23	Tree2	0

Jak widać, wzbogacenie danych o dodatkowe informacje nie ma wpływu na zmianę skuteczności poszczególnych narzędzi analitycznych w konstruowaniu modeli.

Ostatnie dwa zestawienia stanowią ranking uzyskanych modeli uszeregowany według sumy znormalizowanych błędów średniokwadratowych. W odróżnieniu od poprzedniego rankingu, uwzględniają one informację o wszystkich, a nie tylko najlepszych modelach stworzonych dla danego zbioru danych. Ranking przedstawiony

w tabeli 27 oparty jest na znormalizowanych sumach błędów średniokwadratowych wyliczonych na podstawie podzbiorów walidacyjnych.

Tabela 26. Ranking wszystkich metod analitycznych wg liczby zajęcia pierwszego miejsca dla podzbioru danych B (dane wzbogacone)
Źródło: opracowanie własne

L.p.	Narzędzie analityczne	Liczba najlepszych modeli (podzbiór B)
1	PG	6
2	Tree	5
3	Neural	3
4	DmineReg	2
5	Ensmbl	2
6	PLS	2
7	AutoNeural	1
8	LARS	1
9	AutoNeural2	0
10	Boost	0
11	Boost2	0
12	DmineReg2	0
13	DMNeural	0
14	DMNeural2	0
15	Ensmbl2	0
16	LARS2	0
17	MBR	0
18	MBR2	0
19	Neural2	0
20	PLS2	0
21	Reg	0
22	Reg2	0
23	Tree2	0

Z kolei zestawienie znajdujące się w tabeli 28 sporządzono w ten sam sposób, ale oparte jest na podzbiorach treningowych. Jak widać, skumulowana różnica błędu sięga 10%. Natomiast dla obydwu podzbiorów danych uszeregowanie w rankingach (przynajmniej na początkowych miejscach) jest takie samo.

Na zakończenie prezentujemy jeszcze przykładowy wykres obrazujący zależność funkcji dopasowania (minimalizacji błędu średniokwadratowego) od liczby generacji programów w metodzie tworzenia modeli za pomocą programowania genetycznego (rysunek 20). Czerwonym kolorem oznaczony jest wynik najlepszego pojedynczego programu, podczas gdy kolor zielony pokazuje najlepszy wynik rozwiązania zbiorczego.

Tabela 27. Ranking wszystkich metod analitycznych wg suma znormalizowanych błędów obliczonej dla próby walidacyjnej
Źródło: opracowanie własne

L.p.	Model	Suma znormalizowanych błędów (walidacja)
1	Tree	13,90587883
2	PG	14,48214533
3	Neural	18,0497824
4	Ensmbl	18,16807449
5	DmineReg	18,78294511
6	LARS	22,17795715
7	Reg	24,23950957
8	PLS	25,62420136
9	DMNeural	27,23978284
10	Boost	27,76531549
11	MBR2	30,47225519
12	Neural2	31,34130525
13	Tree2	31,34917927
14	AutoNeural	31,37106041
15	Ensmbl2	32,08284908
16	MBR	32,46942853
17	DmineReg2	33,33598569
18	Boost2	35,66328588
19	Reg2	36,79349825
20	LARS2	37,0883756
21	PLS2	37,24142393
22	DMNeural2	37,61626668
23	AutoNeural2	44,96137412

Rysunek 20. Przykładowy wykres obrazujący ewolucję modelu metodą programowania genetycznego
Źródło: oprogramowanie do programowania genetycznego

Tabela 28. Ranking wszystkich metod analitycznych wg suma znormalizowanych błędów obliczonej dla próby treningowej
Źródło: opracowanie własne

L.p.	Model	Suma znormalizowanych błędów (trening)
1	Tree	12,79551593
2	PG	14,08377327
3	Neural	16,59044649
4	Ensmbl	16,99692072
5	DmineReg	17,54329682
6	LARS	21,22148459
7	Reg	22,11998726
8	PLS	24,31110898
9	DMNeural	25,94245684
10	Boost	26,0086734
11	MBR2	27,4790531
12	Tree2	27,54750711
13	DmineReg2	27,66539588
14	Neural2	28,68228791
15	MBR	29,75349386
16	Ensmbl2	29,84102391
17	AutoNeural	29,98527432
18	Boost2	33,38611637
19	Reg2	34,90339267
20	LARS2	34,92601823
21	DMNeural2	35,3419063
22	PLS2	35,53701051
23	AutoNeural2	43,36476126

7.6 Scenariusz wykorzystanie narzędzia do badań

Wskazać można szereg potencjalnych zastosowań dla prezentowanej metody. Do najciekawszych należy zaliczyć: monitorowanie rynku, zasilanie portali ze zbiorczymi ofertami, tworzenie alternatywnego modelu interoperacyjności, czy wreszcie cele badawczo-naukowe. Wszystkie z powyższych zastosowań za wspólny rdzeń mają zaproponowaną metodę. Dopiero po otrzymaniu modeli wyliczenia składki różnić się będą dalszym postępowaniem.

Najciekawszym zastosowaniem wydaje się możliwość automatycznego monitorowania rynku. Przez porównywanie modeli w czasie otrzymać można zarówno obraz zmian w skali całego rynku, jak również w skali konkretnych firm. Doświadczenie z przeprowadzonego eksperymentu wskazuje, że modele i wyliczenia składki ulegają zmianą istotnie częściej niż początkowo było to przewidywane. Dodatkowo takie ele-

menty marketingowe jak czasowe promocje mogą czynić obraz analizowanego rynku jeszcze bardziej dynamicznym.

Możliwe powinno stać się także analizowanie polityki podmiotów w zakresie optymalizacji własnej struktury produktowej oraz reakcji tych podmiotów na oddziaływanie otoczenia. Pod względem analitycznym to zastosowanie jest zbliżone do celów badawczo-naukowych, przy czym w przypadku tej ostatniej grupy zastosowań celem jest nie tyle sam monitoring i analiza, co dalsze przetworzenie i refleksja nad otrzymaną wiedzą.

Dwa środkowe zastosowania, tj. zasilanie portali ze zbiorczymi ofertami oraz tworzenie alternatywnego modelu interoperacyjności stanowiły, w gruncie rzeczy, bezpośrednią przesłankę rozpoczęcia pracy nad przedstawianym projektem badawczym. Jest to też jeden z powodów, dla których takie portale porównujące ofertę same nie były przedmiotem analizy – zależało nam raczej na dotarciu do źródeł WWW podmiotów bezpośrednio oferujących produkt ubezpieczeniowy. Modele stworzone za pomocą przedstawionej metody mogą zostać bezpośrednio zintegrowane z infrastrukturą portali ze zbiorczymi ofertami[263]. Dodatkowo zastosowanie wspólnej platformy konceptualnej w postaci użytej ontologii mogłoby przynieść w dłuższym okresie dodatkowe korzyści dla podmiotów odpowiedzialnych za techniczne wsparcie w takich portalach. Jest to zarazem jeden z możliwych przyczynków do rozwoju alternatywnego modelu interoperacyjności.

[263] Zależy to jeszcze od ostatecznie otrzymanej formy modelu.

8 Wyniki i konkluzje

Na podstawie przeprowadzonych badań i zaprezentowanych rezultatów trudno jest jednoznacznie wskazać narzędzie analityczne, które będzie niekwestionowanym liderem jeśli chodzi o tworzenie wtórnych modeli składki w ramach prezentowanej metody. Na pewno do najbardziej obiecujących zaliczyć można programowanie genetyczne oraz drzewa decyzyjne, przy czym pozycja drugiego narzędzia jest dość nieoczekiwana. W szczególności zaś zaskoczeniem jest fakt, że drzewa decyzyjne osiągnęły lepszy rezultat niż sieci neuronowe.

Oczywiście analizując otrzymane rezultaty należy uwzględnić fakt, iż praktycznie wszystkie użyte narzędzia analityczne generują modele w oparciu o zadane warunki i parametry. Warunki te mogą mieć istotny wpływ na kształt oraz jakość końcowego wyniku. Podobny wpływ ma też przyjęte kryterium stopu w przypadku narzędzi budujących modele w procesie optymalizacji. Dodatkowo przeprowadzone próby wskazują, że w przypadku niektórych metod zamiana parametrów może przyczynić się do poprawy jakości utworzonych modeli. Sytuacja ta dotyczy bardziej takich narzędzi analitycznych, jak sieci neuronowe oraz drzewa decyzyjne. Nie udało się natomiast uzyskać istotnej poprawy wyników w przypadku programowania genetycznego. Tak czy inaczej osiągnięte poziomy miar oceny jakości modeli należy uznać za bardzo zachęcające.

Interesujący jest także powód, dla którego akurat drzewa decyzyjne zajęły wysoką pozycję w rankingach otrzymanych modeli. Prawdopodobnie wynika to z natury tego narzędzia analitycznego i jego szczególnej zdolności do kodowania wiedzy o zmiennych dyskretnych (nominalnych) oraz przedziałowych (interwałowych). Być może w zebranych danych mają one większy wpływ na różnicowanie wysokości składki, mimo że zmiennych typu nominalnego było pod względem liczby nieco mniej (tabela 15).

Pewną zaletą drzew decyzyjnych w stosunku do np. sieci neuronowych jest to, że mogą one przyjąć formę algorytmiczną lub być prezentowane w formie diagramu (przynajmniej, jeśli nie są nazbyt rozbudowane). W tym przypadku jednak, jak już

to wcześniej podkreślono, metoda programowania genetycznego cechuje się podobną zaletą.

Z kolei spoglądając na rezultaty wcześniejszych etapów eksperymentu należy pozytywnie ocenić poziom błędów na etapie ekstrakcji, choć początkowo zakładano, że poziom ten będzie miał bardziej marginalny charakter. Niewątpliwie dla uzyskania jak najlepszych i kompletnych modeli wyceny składki istotna jest wielkość zbioru wyekstrahowanych danych. W wypadku naszego eksperymentu liczbę zbiorów oceniamy bardzo dobrze, natomiast wielkość tych zbiorów jest naszym zdaniem na poziomie zadowalającym. Zapewne interesujące były też wyniki eksperymentu polegającego na utworzeniu dodatkowych zbiorów danych wzbogaconych o informację zewnętrzną. Ten element eksperymentu również można uznać za udany, tzn. zgodnie z oczekiwaniami, zbiory wzbogacone dały możliwość stworzenia nieco lepszych modeli wyceny składki[264]. Różnica nie była co prawda bardzo zauważalna, ale być może wynikało to z dwóch zasadniczych powodów: po pierwsze, dobrych wyników modeli uzyskanych już ze zwykłych (niewzbogaconych) zbiorów danych, po drugie, niewątpliwy wpływ na poprawę miar jakości ma dobór i jakość informacji dodatkowej. Ponieważ nie był to dla nas pierwszoplanowy cel badawczy, więc zapewne w zakresie doboru takiej informacji można by uzyskać lepszy rezultat.

Celem pracy było wykazanie tezy. W rozdziale 5 i 6 przedstawiono wszystkie istotne elementy metody ekstrakcji składki ze źródeł internetowych. W rozdziale 7 zaprezentowano wyniki przeprowadzonego eksperymentu opartego na metodzie. Pozytywny rezultat eksperymentu pozwala uznać, że metoda jest skuteczna, a jej realizacja daje zamierzone wyniki. Metoda i rozwinięte narzędzia są zarazem dostatecznie ogólne i pozwalają na ekstrakcję modeli składki ze zróżnicowanej populacji źródeł internetowych, co zostało zademonstrowane. Niniejszym potwierdza to postawioną na początku pracy tezę.

[264] Przewaga nie była co prawda bardzo duża, jeżeli porównać tylko modele najlepsze. W takim przypadku zależność taka wystąpiła w 12 na 22 przypadki. Rezultat taki wynika jednak z faktu, że przy takim porównaniu nie zawsze porównujemy modele wygenerowane przez to samo narzędzie analityczne.

Aneks A – Język opisu procesu ekstrakcji

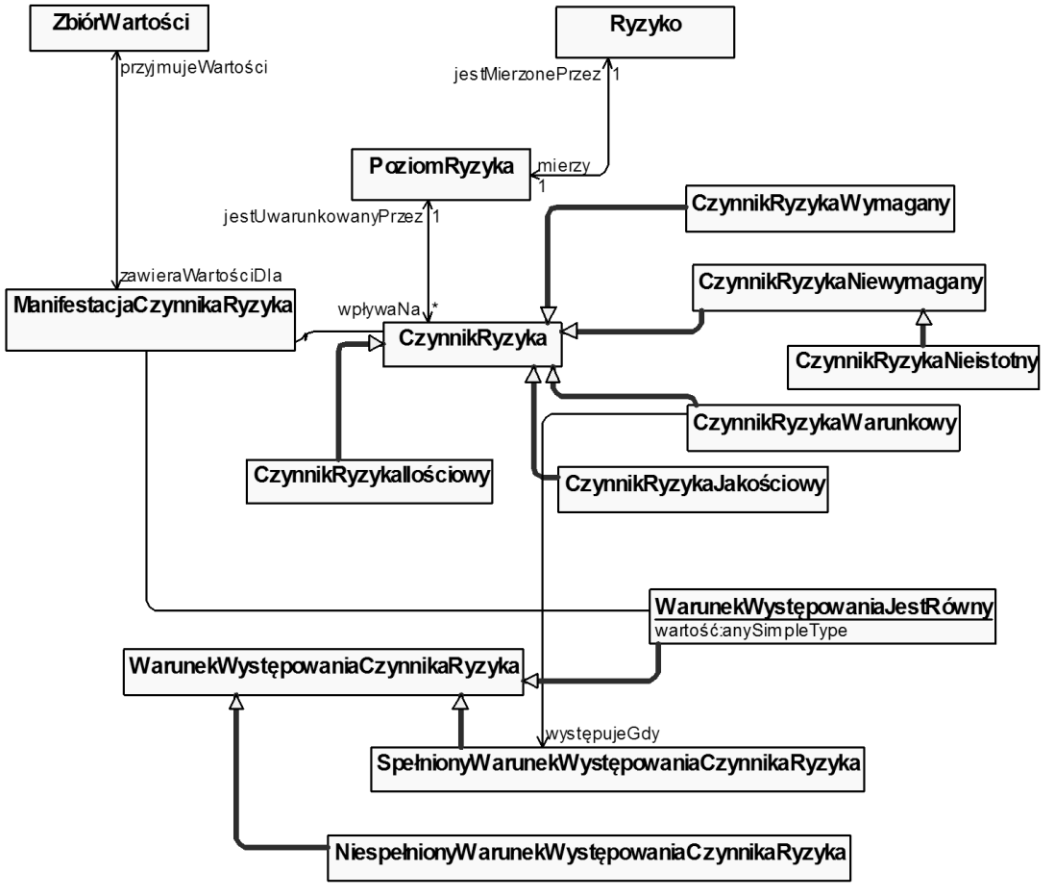
```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" ↵ xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xmlns:xs="http://www.w3.org/2001/XMLSchema" ↵ attributeFormDefault="unqualified" elementFormDefault="qualified">
  <xsd:element name="graph">
    <xsd:complexType>
      <xsd:sequence>
        <xsd:element maxOccurs="unbounded" name="vertex">
          <xsd:complexType>
            <xsd:sequence>
              <xsd:element name="EngineNode">
                <xsd:complexType>
                  <xsd:sequence>
                    <xsd:element name="Breakpoint" type="xsd:boolean" />
                    <xsd:element minOccurs="0" name="Expressions">
                      <xsd:complexType>
                        <xsd:sequence>
                          <xsd:element maxOccurs="unbounded" name="ArrayOfString">
                            <xsd:complexType>
                              <xsd:sequence>
                                <xsd:element name="string" type="xsd:string" />
                              </xsd:sequence>
                            </xsd:complexType>
                          </xsd:element>
                        </xsd:sequence>
                      </xsd:complexType>
                    </xsd:element>
                    <xsd:element minOccurs="0" name="ExtractionItemPattern">
                      <xsd:complexType>
                        <xsd:sequence>
                          <xsd:element name="Visibility" />
                        </xsd:sequence>
                      </xsd:complexType>
                    </xsd:element>
                    <xsd:element minOccurs="0" name="DoGlueData" type="xsd:boolean" />
                    <xsd:element minOccurs="0" name="GlueSeparator" type="xsd:string" />
                    <xsd:element minOccurs="0" name="DoUseProxy" type="xsd:boolean" />
                    <xsd:element minOccurs="0" name="SiteUrl" type="xsd:string" />
                    <xsd:element minOccurs="0" name="Gets" />
                    <xsd:element minOccurs="0" name="Posts" />
                    <xsd:element minOccurs="0" name="DefaultIntertaskTimer" type="AbstractTimerNode">
                    </xsd:element>
                    <xsd:element minOccurs="0" name="CustomReloadtaskTimer" type="AbstractTimerNode">
                    </xsd:element>
                    <xsd:element minOccurs="0" name="CustomAjaxTaskTimer" type="AbstractTimerNode">
                    </xsd:element>
                    <xsd:element minOccurs="0" name="Tasks">
                      <xsd:complexType>
                        <xsd:sequence>
                          <xsd:element maxOccurs="unbounded" name="NodeOfAutomationTask">
                            <xsd:complexType>
                              <xsd:sequence>
                                <xsd:element name="Children" />
                                <xsd:element name="Data">
                                  <xsd:complexType>
                                    <xsd:sequence>
                                      <xsd:element name="Annotations">
                                        <xsd:complexType>
                                          <xsd:sequence minOccurs="0">
                                            <xsd:element name="anyType" type="xsd:string" />
                                          </xsd:sequence>
                                        </xsd:complexType>
                                      </xsd:element>
                                      <xsd:element name="IsDynamic" type="xsd:boolean" />
                                      <xsd:element minOccurs="0" name="Breakpoint" type="xsd:boolean" />
                                      <xsd:element name="Description">
                                        <xsd:complexType>
                                          <xsd:sequence>
                                            <xsd:element name="Key" type="xsd:string" />
                                            <xsd:element name="Value">
                                              <xsd:complexType>
                                                <xsd:sequence>
                                                  <xsd:element maxOccurs="unbounded" name="anyType">
                                                  </xsd:element>
                                                </xsd:sequence>
                                              </xsd:complexType>
                                            </xsd:element>
                                          </xsd:sequence>
                                        </xsd:complexType>
                                      </xsd:element>
```

```
                                </xsd:sequence>
                              </xsd:complexType>
                            </xsd:element>
                          </xsd:sequence>
                        </xsd:complexType>
                      </xsd:element>
                    </xsd:sequence>
                  </xsd:complexType>
                </xsd:element>
                <xsd:element minOccurs="0" name="OutFilePath" type="xsd:string" />
                <xsd:element minOccurs="0" name="OutFileName" type="xsd:string" />
                <xsd:element minOccurs="0" name="IgnoredExtractorIds" />
                <xsd:element minOccurs="0" name="DoAttachHeader" type="xsd:boolean" />
                <xsd:element minOccurs="0" name="DoCombineExtractors" type="xsd:boolean" />
                <xsd:element minOccurs="0" name="Separator" type="xsd:unsignedByte" />
                <xsd:element minOccurs="0" name="Properties">
                  <xsd:complexType>
                    <xsd:sequence minOccurs="0">
                      <xsd:element maxOccurs="unbounded" name="SerializableKeyValuePairOfStringObject">
                        <xsd:complexType>
                          <xsd:sequence>
                            <xsd:element name="Key" type="xsd:string" />
                            <xsd:element name="Value" type="AbstractIterableProperty">
                            </xsd:element>
                          </xsd:sequence>
                        </xsd:complexType>
                      </xsd:element>
                    </xsd:sequence>
                  </xsd:complexType>
                </xsd:element>
                <xsd:element minOccurs="0" name="RequestedInternetExplorerVersion" type="xsd:string" />
              </xsd:sequence>
              <xsd:attribute name="ID" type="xsd:string" use="required" />
            </xsd:complexType>
          </xsd:element>
        </xsd:sequence>
        <xsd:attribute name="id" type="xsd:string" use="required" />
        <xsd:attribute name="type" type="xsd:string" use="required" />
      </xsd:complexType>
    </xsd:element>
    <xsd:element maxOccurs="unbounded" name="edge">
      <xsd:complexType>
        <xsd:attribute name="id" type="xsd:string" use="required" />
        <xsd:attribute name="source" type="xsd:string" use="required" />
        <xsd:attribute name="target" type="xsd:string" use="required" />
      </xsd:complexType>
    </xsd:element>
  </xsd:sequence>
 </xsd:complexType>
</xsd:element>
<xsd:complexType name="AbstractTimerNode">
 <xsd:sequence>
  <xsd:element name="Breakpoint" type="xsd:boolean" />
 </xsd:sequence>
 <xsd:attribute name="ID" type="xsd:string" use="required" />
</xsd:complexType>
<xsd:complexType name="StaticTimerNode">
 <xs:complexContent>
  <xsd:extension base="AbstractTimerNode">
   <xsd:sequence>
    <xsd:element name="DoUseProxy" type="xsd:boolean" />
    <xsd:element name="Gets" />
    <xsd:element name="Posts" />
    <xsd:element name="SuspensionTime" type="xsd:unsignedShort" />
   </xsd:sequence>
  </xsd:extension>
 </xs:complexContent>
</xsd:complexType>
<xsd:complexType name="ReloadTimerNode">
 <xs:complexContent>
  <xsd:extension base="AbstractTimerNode">
   <xsd:sequence>
    <xsd:element name="DoUseProxy" type="xsd:boolean" />
    <xsd:element name="Gets" />
    <xsd:element name="Posts" />
    <xsd:element minOccurs="0" name="SuspensionTime" type="xsd:unsignedShort" />
   </xsd:sequence>
  </xsd:extension>
 </xs:complexContent>
</xsd:complexType>
<xsd:complexType name="LoadAjaxTimerNode">
 <xs:complexContent>
  <xsd:extension base="AbstractTimerNode">
   <xsd:sequence>
    <xsd:element name="DoUseProxy" type="xsd:boolean" />
```

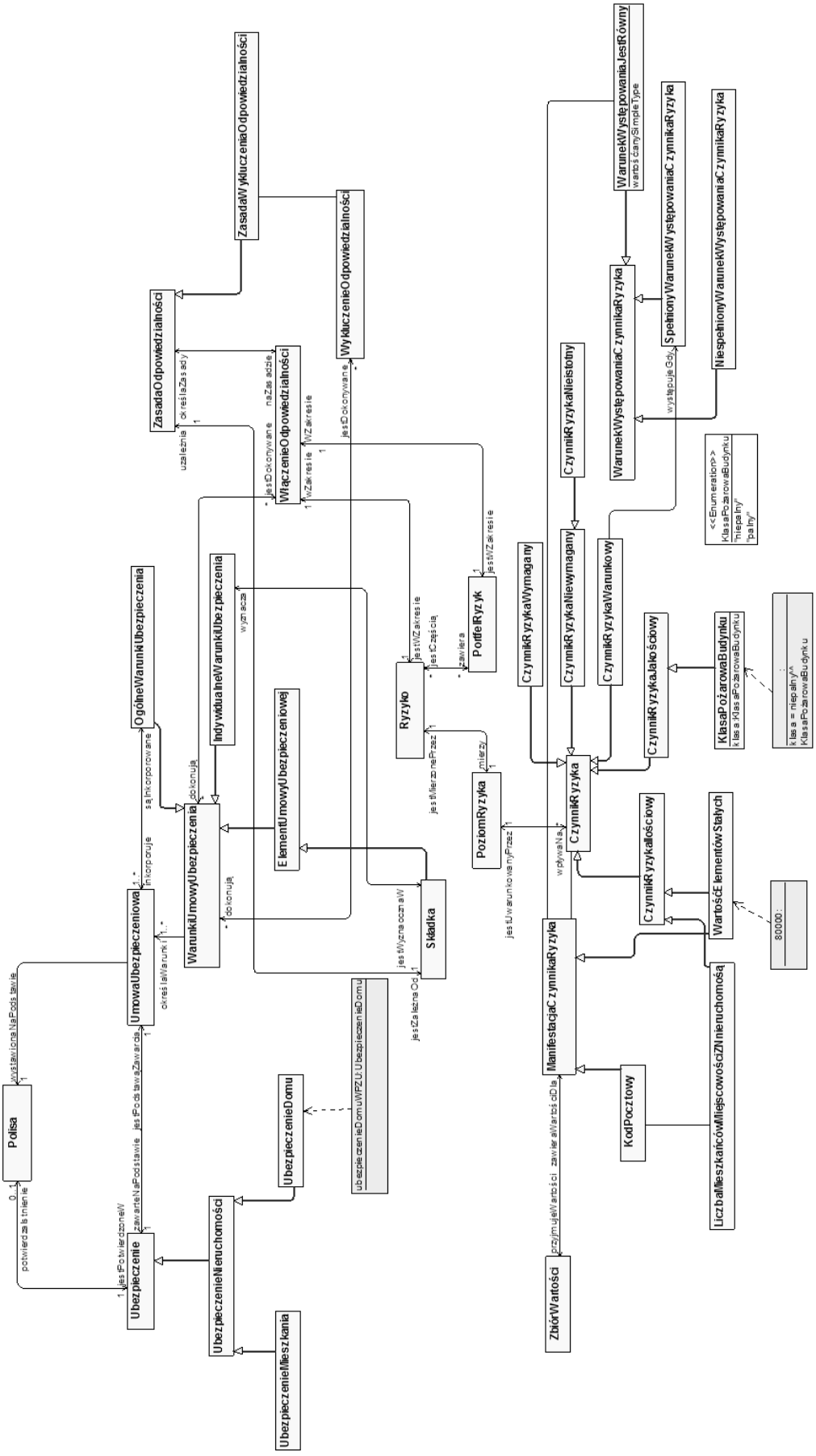
```
        <xsd:element name="Gets" />
        <xsd:element name="Posts" />
        <xsd:element name="MaximumIntertaskPause" type="xsd:unsignedShort" />
        <xsd:element name="LoadedResponsesNumber" type="xsd:unsignedByte" />
        <xsd:element name="WaitUrl" type="xsd:string" />
      </xsd:sequence>
    </xsd:extension>
  </xs:complexContent>
</xsd:complexType>
<xsd:complexType name="XPathExpression">
  <xsd:sequence minOccurs="0">
    <xsd:element name="Expression" type="xsd:string" />
  </xsd:sequence>
</xsd:complexType>
<xsd:complexType name="AbstractIterableProperty">
  <xsd:sequence>
    <xsd:choice maxOccurs="unbounded">
      <xsd:element name="Annotations">
        <xsd:complexType>
          <xsd:sequence minOccurs="0">
            <xsd:element maxOccurs="unbounded" name="anyType" type="xsd:string" />
          </xsd:sequence>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Id" type="xsd:string" />
      <xsd:element name="IsMetaProperty" type="xsd:boolean" />
      <xsd:element name="FormItemReflectedXPathQuery" type="xsd:string" />
      <xsd:element name="TextualValues">
        <xsd:complexType>
          <xsd:sequence minOccurs="0">
            <xsd:element maxOccurs="unbounded" name="string" type="xsd:string" />
          </xsd:sequence>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="OriginalMinimum" type="xsd:string" />
      <xsd:element name="OriginalMaximum" type="xsd:string" />
      <xsd:element name="Format" type="xsd:string" />
      <xsd:element name="PropertyIds">
        <xsd:complexType>
          <xsd:sequence>
            <xsd:element maxOccurs="unbounded" name="SerializableKeyValuePairOfStringString">
              <xsd:complexType>
                <xsd:sequence>
                  <xsd:element name="Key" type="xsd:string" />
                  <xsd:element name="Value" type="xsd:string" />
                </xsd:sequence>
              </xsd:complexType>
            </xsd:element>
          </xsd:sequence>
        </xsd:complexType>
      </xsd:element>
      <xsd:element name="Minimum" />
      <xsd:element name="Step" type="xsd:string" />
      <xsd:element name="Maximum" />
      <xsd:element name="ConditionalValues">
        <xsd:complexType>
          <xsd:sequence>
            <xsd:element maxOccurs="unbounded" name="NodeOfConditionalValue">
              <xsd:complexType>
                <xsd:sequence>
                  <xsd:element name="Children">
                    <xsd:complexType>
                      <xsd:sequence>
                        <xsd:element maxOccurs="unbounded" name="NodeOfConditionalValue">
                          <xsd:complexType>
                            <xsd:sequence>
                              <xsd:element name="Children" />
                              <xsd:element name="Data">
                                <xsd:complexType>
                                  <xsd:sequence>
                                    <xsd:element name="Value" type="xsd:unsignedShort" />
                                  </xsd:sequence>
                                </xsd:complexType>
                              </xsd:element>
                            </xsd:sequence>
                          </xsd:complexType>
                        </xsd:element>
                      </xsd:sequence>
                    </xsd:complexType>
                  </xsd:element>
                  <xsd:element name="Data">
                    <xsd:complexType>
                      <xsd:sequence>
                        <xsd:element name="Condition">
                          <xsd:complexType>
```

```
                        <xsd:sequence>
                          <xsd:element name="PropertyName" type="xsd:string" />
                          <xsd:element name="Relation" type="xsd:string" />
                          <xsd:element name="Value" />
                        </xsd:sequence>
                      </xsd:complexType>
                    </xsd:element>
                  </xsd:sequence>
                </xsd:complexType>
              </xsd:element>
            </xsd:sequence>
          </xsd:complexType>
        </xsd:element>
        <xsd:element name="Current" type="xsd:string" />
      </xsd:choice>
    </xsd:sequence>
  </xsd:complexType>
  <xsd:complexType name="TextualIterableProperty">
    <xs:complexContent>
      <xsd:extension base="AbstractIterableProperty">
        <xsd:sequence>
        </xsd:sequence>
      </xsd:extension>
    </xs:complexContent>
  </xsd:complexType>
  <xsd:complexType name="PropertyFormatterProperty">
    <xs:complexContent>
      <xsd:extension base="AbstractIterableProperty">
        <xsd:sequence>
        </xsd:sequence>
      </xsd:extension>
    </xs:complexContent>
  </xsd:complexType>
  <xsd:complexType name="DateTimeIterableProperty">
    <xs:complexContent>
      <xsd:extension base="AbstractIterableProperty">
        <xsd:sequence>
        </xsd:sequence>
      </xsd:extension>
    </xs:complexContent>
  </xsd:complexType>
  <xsd:complexType name="NumericRandomGeneratedProperty">
    <xs:complexContent>
      <xsd:extension base="AbstractIterableProperty">
        <xsd:sequence>
        </xsd:sequence>
      </xsd:extension>
    </xs:complexContent>
  </xsd:complexType>
  <xsd:complexType name="TextualRandomValuedProperty">
    <xs:complexContent>
      <xsd:extension base="AbstractIterableProperty">
        <xsd:sequence>
        </xsd:sequence>
      </xsd:extension>
    </xs:complexContent>
  </xsd:complexType>
  <xsd:complexType name="ConditionalTextualIterableProperty">
    <xs:complexContent>
      <xsd:extension base="AbstractIterableProperty">
        <xsd:sequence>
        </xsd:sequence>
      </xsd:extension>
    </xs:complexContent>
  </xsd:complexType>
</xs:schema>
```

Aneks B – Ontologia

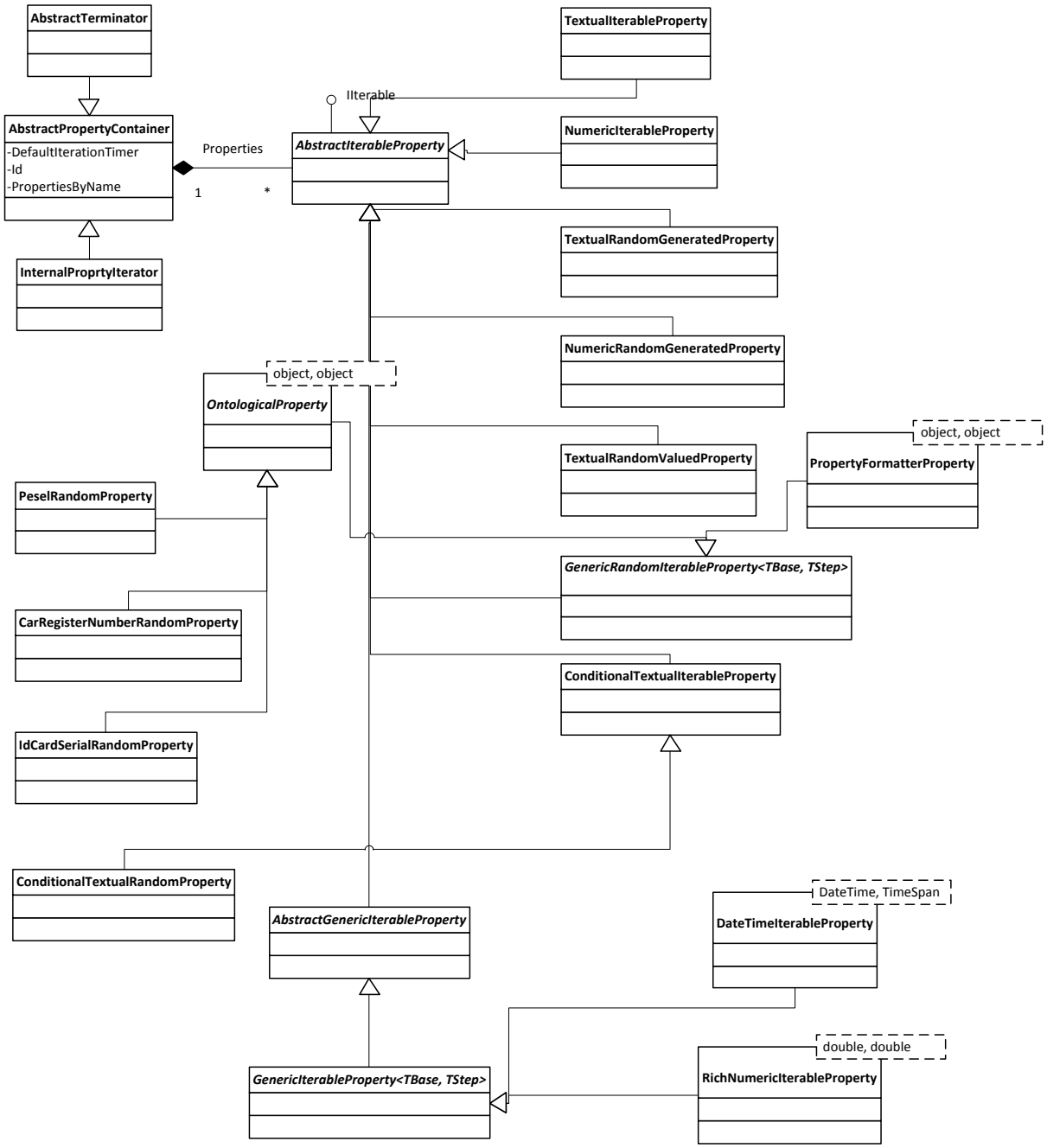


Rysunek 21. Model UML opisujący czynniki ryzyka
Źródło: opracowanie własne

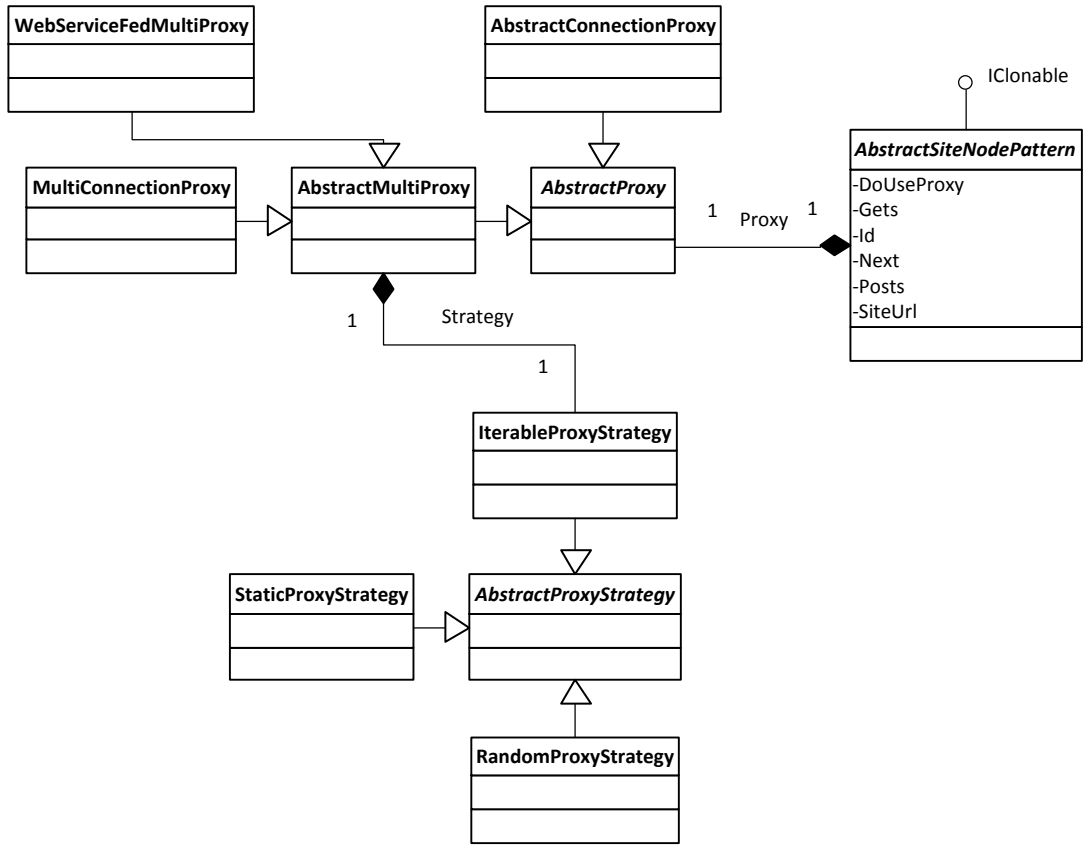


Rysunek 22. Model domeny ubezpieczeń
Źródło: opracowanie własne

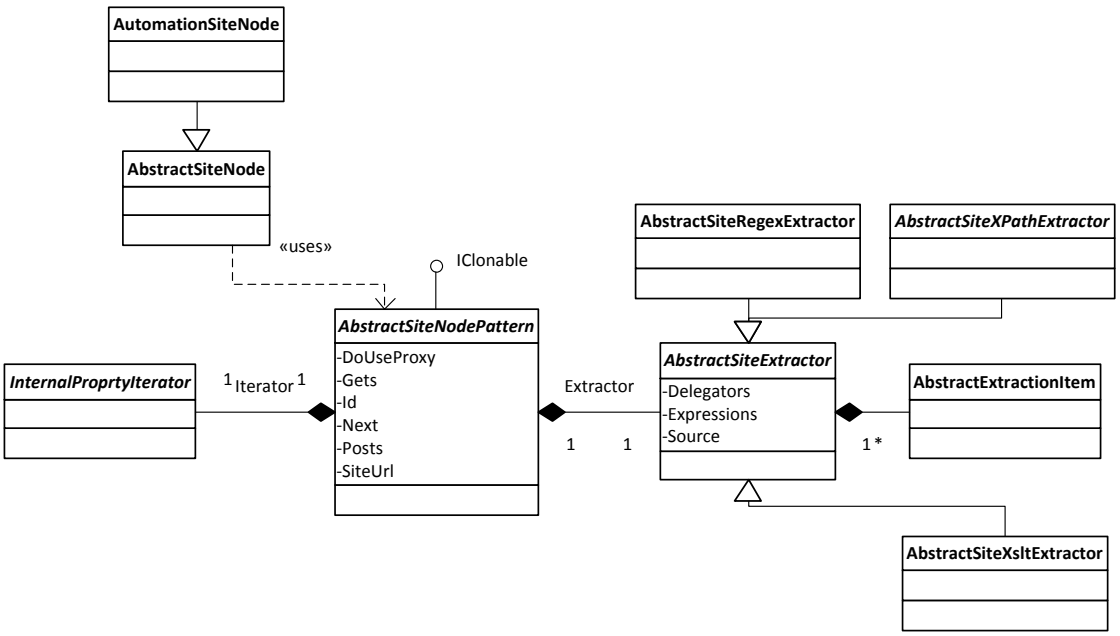
Aneks C – Metoda ekstrakcji – schematy UML



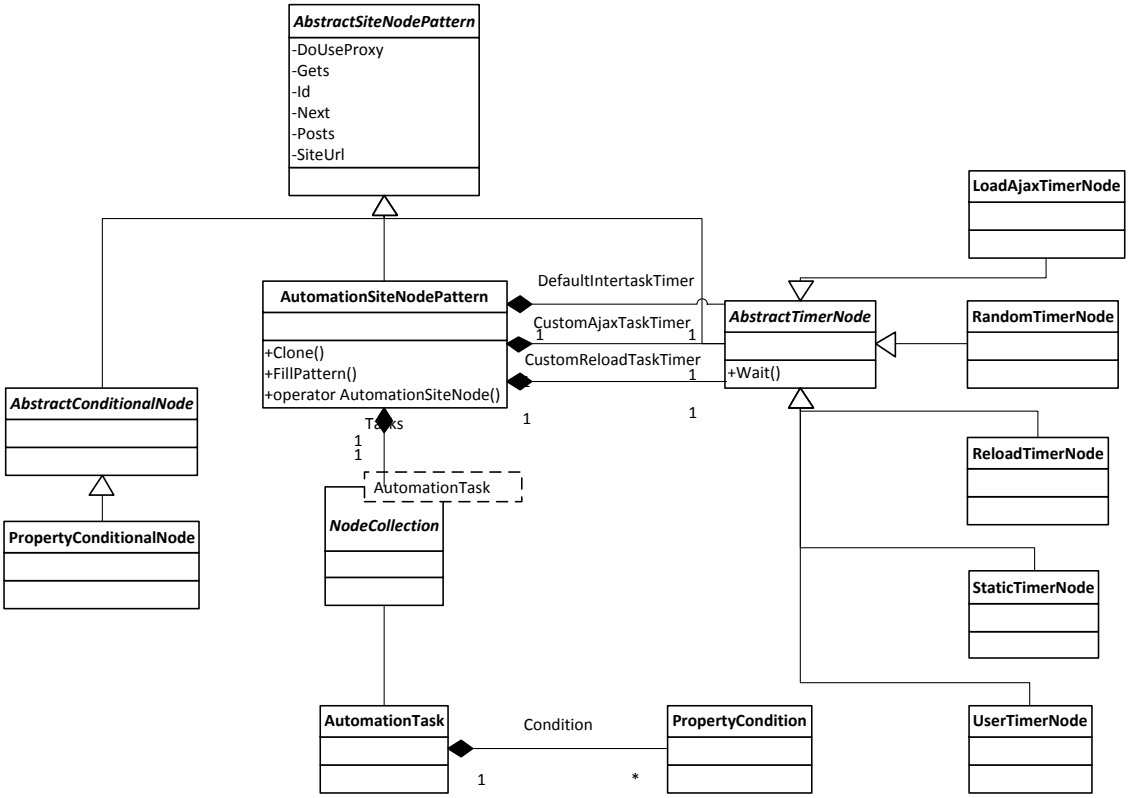
Rysunek 23. Diagram struktury statycznej klas właściwości
Źródło: opracowanie własne



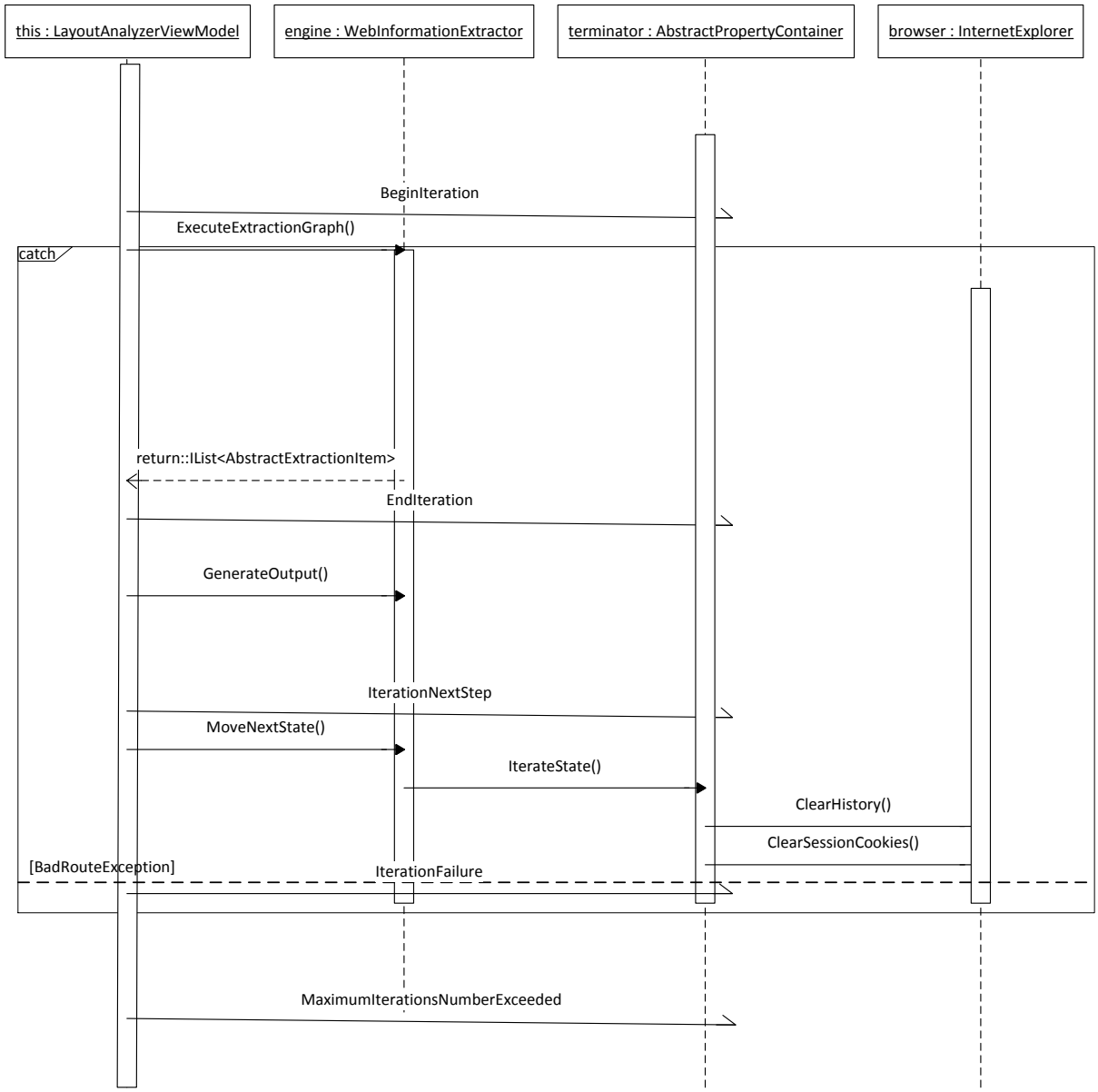
Rysunek 24. Diagram struktury statycznej klas proxy
Źródło: opracowanie własne



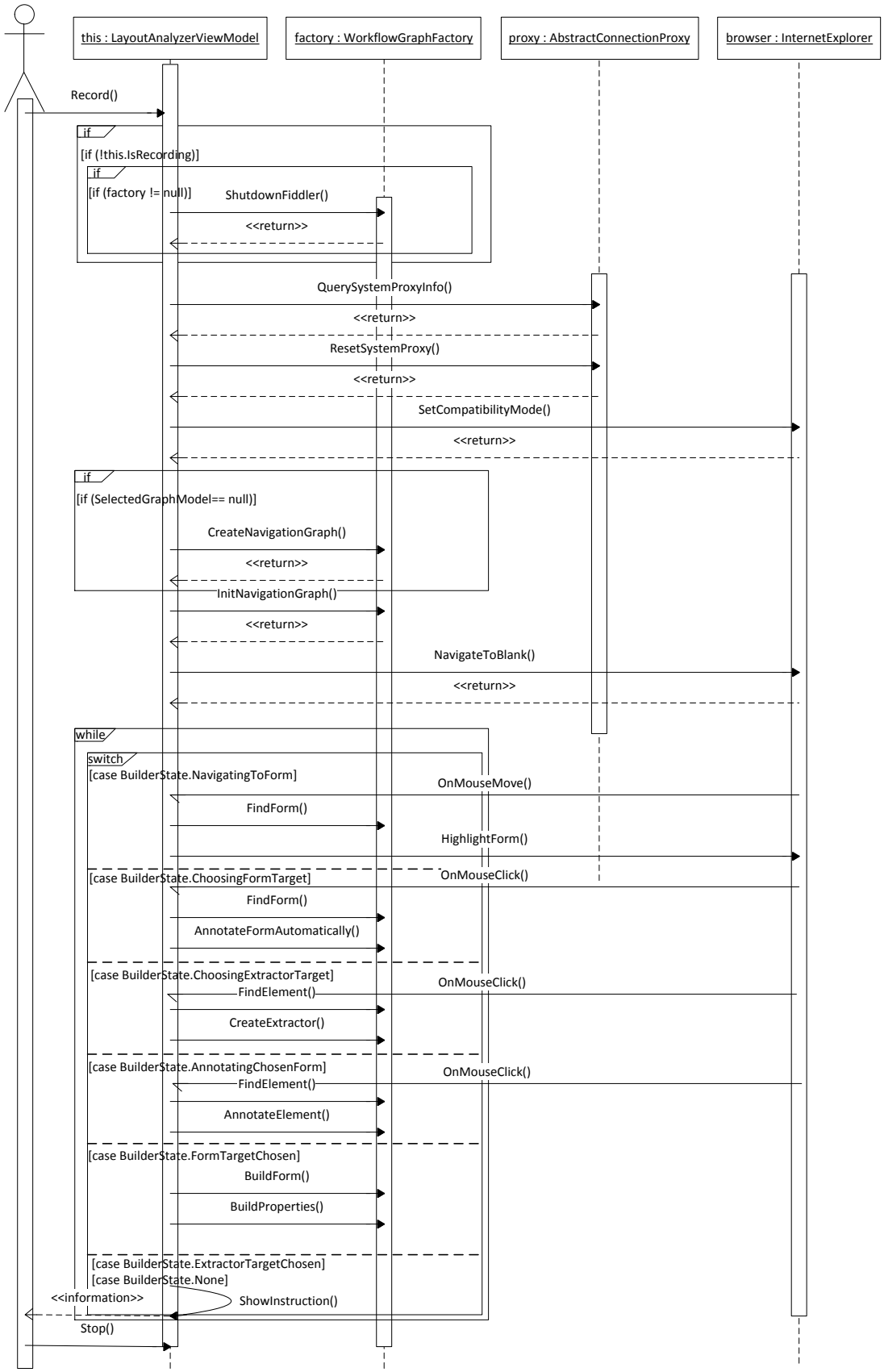
Rysunek 25. Diagram struktury statycznej klas wzorca podstrony oraz ekstraktora
Źródło: opracowanie własne



Rysunek 26. Diagram struktury statycznej klas mierników czasu
Źródło: opracowanie własne



Rysunek 27. Diagram sekwencji nawigacji po źródle webowym
Źródło: opracowanie własne



Rysunek 28. Diagram sekwencji wsparcia budowy grafu
Źródło: opracowanie własne

Bibliografia

[Abramowicz2011] W. Abramowicz, P. Stolarski, K. Węcel, *Ontologie jako narzędzie budowy modeli w ubezpieczeniowych systemach informacyjnych – modelowanie ryzyka oraz produktów*, Wiadomości Ubezpieczeniowe, PIU nr 1, s. 117-137, 2011.

[Adelberg1998] B. Adelberg, *Nodose-a tool for semi-automatically extracting structured and semistructured data from text documents*, 1998 ACM SIGMOD International Conference on Management of Data, s. 283–294, 1998.

[Ahmed2010] Z. H. Ahmed, *Genetic Algorithm for the Traveling Salesman Problem Using Sequential Constructive Crossover Operator*, International Journal of Biometric and Bioinformatics 3/6, 2010.

[Alvarez2007] M. Alvarez, J. Raposo, A. Pan, F. Cacheda, F. Bellas, V. Carneiro, *Deepbot: A focused crawler for accessing hidden web content*, 3rd international workshop on Data engineering issues in E-commerce and services, s. 18–25, 2007.

[Anderson2007] D. Anderson, S. Feldblum, C. Modlin, D. Schirmacher, E. Schirmacher, N. Thandi, *The Practitioner's Guide to Generalized Linear Models*, Casualty Actuarial Society Study Note, 2007.

[Arabas2001] J. Arabas, *Wykłady z algorytmów ewolucyjnych*, WNT, Warszawa, 2001.

[Arasu2005] A. Arasu, H. Garcia-Molina, *Extracting structured data from web pages*, 2003 ACM SIGMOD International Conference on Management of Data, s. 337–348, 2003.

[Arocena1998] G. O. Arocena, A. O. Mendelzon, *Weboql: Restructuring documents, databases, and webs*, 14th International Conference on Data Ingeneering, 1998.

[Ashish1997] N. Ashish, C. A. Knoblock, *Semi-automatic wrapper generation for internet information sources*, w: 2nd International Conference on Cooperative Information Systems, s. 160-169, 1997.

[Azavant1999] F. Azavant, A. Sahuguet. *Bulding light-weight wrappers for legacy web data sources using w4f*, 25th International Conference on Very Large Data Bases, 1999.

[ASOP2004] *Actuarial Standard of Practice No. 23, Data Quality*, Actuarial Standards Board of the American Academy of Actuaries, 2004.

[Baumgartner2001] R. Baumgartner, S. Flesca, G. Gottlob, *Visual web information extraction with lixto*, w: 27th International Conference on Very Large Data Bases, s. 119–1128, 2001.

[Beckett2004] D. Beckett, B. McBride, *RDF/XML syntax specification*, 2004.

[Bergman2001] M. K. Bergman, *The deep web: Surfacing hidden value*, The Journal of Electronic Publishing, 7(1), 2001.

[Berners-Lee2000] T. Berners-Lee, *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web*, San Francisco, Harper. ISBN 9780062515872, 2000.

[Berry2000] M. J. A. Berry, G. S. Linoff, *Mastering Data Mining. The Art and Science of Customer Relationship Management*, Wiley Computer Publishing, New York, s. 494, 2000.

[Błaszczyszyn2004] B. Błaszczyszyn, T. Rolski, *Podstawy matematyki ubezpieczeń na życie*, WNT, Warszawa 2004.

[Bramer2007] M. Bramer, *Principles of Data Mining*, Springer, London, 2007.

[Breiman1984] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone., *Classification and Regression Trees*, Wadsworth and Brooks, Pacific Grove, California, 1984.

[CEPiK] http://www.cepik.gov.pl/, odczytano 25 maja 2015 r.

[Chang2006] Ch, Chang, M. Kayed, M. R. Girgis, K. Shaalan, *A survey of web information extraction systems*, IEEE Transactions on Knowledge and Data Engineering, 18(10), s. 1411–1428, 2006.

[Chawathe1994] S. Chawathe, Y. Papakonstantinou, J. D. Ullman, H. Garcia-Molina, K. Ireland, J. Hammer, J. Widom, *The TSIMMIS project: Integration of heterogeneous information sources*, w: 10th Meeting of the Information Processing Society of Japan, s. 7–18, 1994.

[Childs1980] D. Childs, R. A. Currie, *Expense Allocation in Insurance Ratemaking*, w: Pricing Property and Casualty Insurance Products (Casualty Actuarial Society 1980 Discussion Paper Program), s. 32-66, 1980.

[Clarck1999] J. Clark, *XSL transformations (XSLT) version 1.0*, 1999

[Codd1970] E. F. Codd, *A Relational Model of Data for Large Shared Data Banks*, Communications of the ACM 13 (6), s. 377–387, doi:10.1145/362384.362685, 1970.

[Daszkowska1997] M. Daszkowska, *Usługi. Produkcja, rynek, marketing*, WN PWN, Warszawa, 1997.

[DFA2007] *New Version of DFA's Insurance Modeling Technology Adds Multi-Currency Capabilities and Modeling for Businesses with Complex Structures* (16 October), Business Wire oraz *Eckler Ltd. Signs Multi-Year Deal for ADVISE(R) and GEMS(R) Solutions from DFA* (15 December), Business Wire, 2007.

[Dionne1989] G. Dionne, Ch. Vanasse, *A generalization of actuarial automobile insurance rating models. The negative binomial distribution with a regression component*. Actuarial Bulletin 19, s. 199–212, 1989.

[Dionne1992] G. Dionne, Ch. Vanasse, *Automobile Insurance Ratemaking in the Presence of Asymmetrical Information*, Journal of Applied Econometrics Vol. 7, No. 2 (Apr.-Jun., 1992), s. 149-165, Wiley-Blackwell, http://www.jstor.org/stable/2285025, 1992.

[Doorenbos1997] R. B. Doorenbos, O. Etzioni, D. S. Weld, *A scalable comparison-shopping agent for the world-wide web*, w: 1st International Conference on Autonomous Agents, s. 39–48, 1997.

[Favier2006] J. Favier, M. Bouquet, *Europe's eCommerce Forecast: 2006 To 2011*, Forrester Research, Inc., 2006.

[Feldblum2003] S. Feldblum, J. E. Brosius, *The Minimum Bias Procedures: A Practitioner's Guide*, w: Proceedings of Casualty Actuarial Society, vol. 90, s. 591-684, 2003.

[Fielding2000] R. T. Fielding, *Architectural Styles and the Design of Network-based Software Architectures*, praca doktorska, Information and Computer Science, University of California, Irvine, California, USA, 2000

[Finger2001] R. Finger, *Risk Classification*, w: Foundations of Casualty Actuarial Science 4th ed., Arlington, VA: Casualty Actuarial Society, s. 292-301, 2001.

[Flejter2011] D. Flejter, *Semi-Automatic Web Information Extraction*, praca doktorska, UEP, 2011.

[Florescu1997] D. Florescu, D. Koller, A. Halevy, *Using probabilistic information in dataintegration*, w: 23rd International Conference on Very Large Data Bases, s. 216–225, 1997.

[Fraser1957] A. S. Fraser, *Simulation of genetic systems by automatic digital computers*, J. Biol. Sci., 10, s. 484-499, 1957.

[Gerber1997] H. U. Gerber, *Life Insurance Mathematics*, Springer-Verlag (3 wyd.), 1997.

[Gobble2003] E. Gobble, D. Windeler, *Catastrophe modeling: Shifting perceptions*, Canadian Underwriter, 70(6), s. 38, 2003.

[Haller1998] M. Haller, *Produkt- Und Sortimentsgestaltung*, w: Handswoerterbuch der Versicherung, D. Farny, E. Helten, P. Koch, R. Schmidt, Versicherungswirtschaft, Karlsruhe, 1998.

[Handschke2000] J. Handschke, *Gospodarcze i społeczne znaczenie ubezpieczeń gospodarczych*, w: Vademecum ubezpieczeń gospodarczych, red. T. Sangowski, SAGA Printing, Poznań 2000.

[Hespos1965] R. F. Hespos, P. A. Strassmann, *Stochastic decision trees for the analysis of investment decisions*, Management Science 11/10, s. B244-B259, 1965.

[Hastie2009] T. Hastie, R. Tibshirani, J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer Science & Business Media, ISBN 9780387848587, 2009.

[Hevner2004] A. R. Hevner, S. T. March, J. Park, S. Ram, *Design science in information systems research*, MIS Quarterly, 28(1), s. 75–105, 2004.

[Himmeröder1997] R. Himmeröder, G. Lausen, B. Ludäscher, Ch. Schlepphorst, *On a declarative semantics for web queries*, w: 5th International Conference on Deductive and Object-Oriented Databases, s. 386–398, 1997.

[Hsu1998] Ch. Hsu, M. Dung, *Generating finite-state transducers for semi-structured data extraction from the web*, Information Systems, 23(9), s. 521–538, 1998.

[Iopus2012] http://wiki.imacros.net/, odczytano 22 listopada 2012 r.

[Iskold2007] A. Iskold, *Top-down: A new approach to the semantic web*, 2007.

[Jaworski2010] J. Jaworski, *Teoria i praktyka zarządzania finansami przedsiębiorstw*, CeDeWu, ISBN 9788375562262, 2010.

[Kaczała2006] M. Kaczała, *Internet jako instrument dystrybucji ubezpieczeniowej*, praca doktorska, UEP, 2006.

[Kaczmarek2006] T. Kaczmarek, *Deep Web data integration for company environment analysis*, praca doktorska, UEP, 2006.

[Kaczmarek2010] T. Kaczmarek, D. Zyskowski, A. Walczak, W. Abramowicz, *Information Extraction from Web Pages for the Needs of Expert Finding*, w: Studies in Logic, Grammar and Rethoric, Logic Philosophy and Computer Science, vol 22(35), s. 141-157, ISBN 9788374312738, 2010.

[Kass1980] G. V. Kass, *An exploratory technique for investigating large quantities of categorical data*, Applied Statistics 29(2), s. 119–127, doi:10.2307/2986296, JSTOR 2986296, 1980.

[Kleinberg1999] J. M. Kleinberg, *Authoritative sources in a hyperlinked environment*, Journal of the ACM, 46(5), s. 604–632, 1999.

[Konopnicki1995] D. Konopnicki, O. Shmueli, *W3QS: A query system for the world-wide web*, w: 21st International Conference on Very Large Data Bases, 1995.

[Kosala2000] R. Kosala, H. Blockeel, *Web mining research: a survey*, ACM SIGKDD Explorations Newsletter, 2(1), s. 1–15, 2000.

[Kotler2012] G. Armstrong, P. Kotler, *Marketing: wprowadzenie*, tłum. D. Wąsik, Oficyna Wolters Kluwer business, ISBN 9788326404849, 2012.

[Kowalczyk2006] P. Kowalczyk, E. Poprawska, W. Ronka-Chmielowiec, *Metody aktuarialne*, Warszawa, WN PWN, ISBN 9788301146597, 2006.

[Kowalewski2006] E. Kowalewski, *Prawo Ubezpieczeń Gospodarczych*, wyd. 3, Oficyna Wydawnicza Branta, Bydgoszcz, 2006.

[Koza1994] J. Koza, *Genetic Programming II: Automatic Discovery of Reusable Programs*, MIT Press. ISBN 0262111896, 1994.

[Kuhlins2002] S. Kuhlins, R. Tredwell, *Toolkits for generating wrappers*, Net.ObjectDays 2002: Objects, Components, Architectures, Services and Applications for a Networked World, http://www.netobjectdays.org/, LNCS 2591, 2002.

[Kushmerick1997] N. Kushmerick, *Wrapper induction for information extraction*, praca doktorska, University of Washington, 1997.

[Kushmerick2003] N. Kushmerick, B. Thomas, *Adaptive Information Extraction: Core technologies for Information agents*, Springer, s. 79–103, 2003.

[Lane1985] J. Lane, D. Glennon, *The Estimation of Age/Earnings Profiles in Wrongful Death and Injury Cases*, Journal of Risk and Insurance, 52(4), s. 686-695, 1985.

[Lange1989] K. L. Lange, R. J. A. Little, J. M. G. Taylor, *Robust Statistical Modeling Using the t Distribution*, Journal of the American Statistical Association 84 (408), s. 881–896. doi:10.2307/2290063, JSTOR 2290063, 1989.

[Lausen2007] H. Lausen, T. Haselwanter, *Finding Web Services*, w: Proceedings of the 1st European Semantic Technology Conference (ESTC), 2007.

[Liu2000] L. Liu, C. Pu, W. Han, *XWRAP: An XML-enabled wrapper construction system for web information sources*, w: 16th International Conference on Data Engineering, s. 611–621, 2000.

[Magee1964] J. F. Magee, *Decision trees for decision making*, Harvard Business Review, 1964.

[Masters1993] T. Masters, *Sieci neuronowe w praktyce*, WNT, Warszawa, 1996.

[McCallum2002] A. McCallum, W. W. Cohen, *Information extraction from the world wide web*, tutorial, 2002.

[McCulloch1943] W. S. McCulloch, *A logical calculus of the ideas immanent in nervous activity*, Bulletain of Mathematical Biophysics, nr 5, s. 115-133, 1943.

[McGuinness2004] D. L. McGuinness, F. van Harmelen, *OWL web ontology language overview* (W3C recommendation 10 february 2004), 2004.

[Mendelzon1997] A. O. Mendelzon, G. A. Mihaila, T. Milo, *Querying the world wide web*, International Journal on Digital Libraries, 1(1), s. 54–67, 1997.

[Merialdo2001] P. Merialdo, G. Mecca, V. Crescenzi, *RoadRunner: Towards automatic data extraction from large web sites*, w: 27th International Conference on Very Large Data Bases, s. 109–118, 2001.

[Michalski2004] A. Karmańska, T. Michalski, A. Śliwiński, *Ubezpieczenia gospodarcze: ryzyko i metodologia oceny*, pod red. Tomasza Michalskiego, C. H. Beck, Warszawa, 2004.

[Miller1989] G. F. Miller, P. M. Todd, S. U. Hagde, *Designing neural networks Rusing genetic algorithms*, Proc. of the 3rd Int. Conf. on Genetic Algorithms and Their Applications, Schaffer J.D. (red.), Morgan Kaufmann, s. 379-384, San Mateo, USA, 1989.

[Monkiewicz2000] J. Monkiewicz, L. Gąsiorkiewicz, B. Hadyniak, *Zarządzanie finansami ubezpieczeń*, Poltext, 2000.

[Montana1989] D. Montatna, L. Davis, *Training feedforward neural networks using genetic algorithms*, Proc. of the 11th Int. Conf. on Artificial Intelligence, Morgan Kaufmann, s. 762-767, USA, 1989.

[Mulpuru2011] S. Mulpuru, V. Sehgal, P. F. Evans, D. Roberge, *Forecast: US Online Retail Sales, 2010 to 2015*, Forrester Research, Inc., 2010.

[Muslea1999] I. Muslea, S. Minton, C. A. Knoblock. *A hierarchical approach to wrapper induction*, w: 3rd International Conference on Autonomous Agents, s. 190–197, 1999.

[NetTrack2011] Badanie NetTrack, MillwardBrown SMG/KRC, 2011.

[NetTrack2015] http://www.wirtualnemedia.pl/artykul/23-mln-internautow-w-polsce-29-proc-z-dostepem-mobilnym-najczesciej-korzystaja-z-maili-i-e-zakupow, odczytano 29 czerwca 2015 r.

[Nievergelt1994] Y. Nievergelt, *Total Least Squares: State-of-the-Art Regression in Numerical Analysis*, SIAM Review 36 (2), s. 258–264, doi:10.1137/1036055, JSTOR 2132463, 1994.

[Nonaka1995] I. Nonaka, H. Takeuchi, *The Knowledge-creating Company: How Japanese Companies Create the Dynamics of Innovation*, Oxford University Press, ISBN 9780195092691, 1995.

[Nordin1999] P. Nordin, F. Hoffmann, F. D. Francone, M. Brameier, W. Banzhaf, *AIM-GP and parallelism*, w: Evolutionary Computation, CEC 99, Proceedings of the 1999 Congress on (Vol. 2), IEEE, 1999.

[OECD2011] *OECD Broadband Portal: Yearly penetration increase*, 2011 http://www.oecd.org/document/54/0,3746,en_2649_37441_38690102_1_1_1_37441,00.html, odczytano: 13 marzec 2013 r.

[OIUFG] http://www.ufg.pl/en/web/guest/baza-oc-i-ac, odczytano: 17 luty 2013 r.

[Osowski2000] S. Osowski, *Sieci neuronowe do przetwarzania informacji*, Wydawnictwo PW, Warszawa, 2000.

[Pao1989] Y. Pao, *Adaptive Pattern Recognition and Neural Networks*, Addison-Wesley Publishing Co., Reading, MA, 1989.

[Page1998] L. Page, S. Brin, R. Motwani, T. Winograd, *The PageRank citation ranking: Bringing order to the web*, Technical report, 1998.

[Papazoglou2006] M. P. Papazoglou, P. Ribbers, *e-Business – Organizational and Technical Foundations*, John Wiley & Sons, 2006.

[Quinlan1986] J.R. Quinlan, *Induction of Decision Trees*, Machine Learning, nr 1, Morgan Kaufmann , s. 81-106, 1986.

[Quinlan1996] J. R. Quinlan. *Improved use of continuous attributes in c4.5*, Journal of Artificial Intelligence Research, nr 4: s. 77-90, 1996.

[Riley1985] J. G. Riley, *Competition with Hidden Knowledge*, Journal of Political Economy Vol. 93, nr 5, s. 958-976, 1985.

[Ronka-Chmielowiec2006] P. Kowalczyk, E. Poprawska, W. Ronka-Chmielowiec, *Metody aktuarialne: zastosowania matematyki w ubezpieczeniach*, red. nauk. W. Ronka-Chmielowiec, Wydawnictwo Naukowe PWN, 2006.

[Rosenblatt1958] F. Rosenblatt, *The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain*, Psychological Review 65, s. 386-408, 1958.

[Rothschild1976] M. Rothschild, J. Stiglitz, *Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information*, The Quarterly Journal of Economics, Vol. 90, nr 4, s. 629-649, Oxford University Press, http://www.jstor.org/stable/1885326, 1976.

[Rss2007] *RSS 2.0 specification (version 2.0.10)*, 2007.

[Rumelhart1986] D. E. Rumelhart, J. L. McClelland, the PDP research group, *Parallel distributed processing: Explorations in the microstructure of cognition*, Cambridge, MA: MIT Press, 1986.

[Salam2003] R. Salam, *Estimating the Cost of Commercial Airlines Catastrophes A Stochastic Simulation Approach*, w: The Casualty Actuarial Society Forum Winter 2003 Edition Including the Data Management Call Papers and Ratemaking Discussion Papers, s. 379, 2003.

[Schofield1998] D. Schofield, *Going from a Pure Premium to a Rate*, CAS Study Note, 1998.

[Shestakov2005] D. Shestakov, S. S. Bhowmick, E. Lim, *Deque: querying the deep web*, Data Knowl. Eng., 52(3), s. 273–311, 2005.

[Soderland1997] S. Soderland, *Learning to extract text-based information from the world wide web*, 3rd International Conference on Knowledge Discovery and Data Mining, s. 251–254, 1997.

[Sternik2009] T. Sternik, *Systemy informatyczne obsługujące współczesną działalność ubezpieczeniową. Rodzaje, rola i uwarunkowania*, Wiadomości Ubezpieczeniowe, PIU nr 3, s. 36-55, 2009.

[Stolarski2012] W. Abramowicz, P. Stolarski, K. Węcel, *Ontologie jako narzędzie budowy modeli w ubezpieczeniowych systemach informacyjnych – ekstrakcja wiedzy ubezpieczeniowej ze źródeł internetowych*, Wiadomości Ubezpieczeniowe, PIU nr 2, s. 9-28, 2012.

[Strobl2009] C. Strobl, J. Malley, G. Tutz, *An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests*, Psychological Methods 14 (4), s. 323–348, doi:10.1037/a0016973, 2009.

[Werner2010] G. Werner, C. Modlin, *Basic Ratemaking 4th ed.*, Casualty Actuarial Society, 2010.

[Węcel2011] K. Węcel, W. Abramowicz, P. Stolarski, A. Filipowska, B. Perkowski, *System for Detection of Illegal Drugs E-Trading*, w: Frontiers in Artificial Intelligence and Applications 235, IOS Press, Vienna, Austria, ISBN 9781607509806, 2011.

[Wolny-Dominiak2011] A. Wolny-Dominiak, *Analiza porównawcza modeli mieszanych szacowania stóp taryf w ubezpieczeniach majątkowych z wykorzystaniem kroswalidacji*, Wydawnictwo Uniwersytetu Ekonomicznego, Wrocław, s. 229-237, 2011.

[Tomaszewski2009] T. Tomaszewski, *Przykłady zastosowań koncepcji semantycznej reprezentacji ryzyka w ubezpieczeniowych systemach informacyjnych*, w: Studia Ubezpieczeniowe, Wydaw. AE. s. 232-242, 2009.

[Zhang2008] Q. Zhang, R. S. Segall, *Web Mining: a Survey of Current Research, Techniques, and Software*, International Journal of Information Technology and Decision Making 7(4), s. 683-720, 2008.