WITOLD ABRAMOWICZ
PIOTR STOLARSKI
KRZYSZTOF WĘCEL

Ontologie jako narzędzie budowy modeli w ubezpieczeniowych systemach informacyjnych – ekstrakcja wiedzy ubezpieczeniowej ze źródeł internetowych

Ekstrakcja modeli wyceny ubezpieczeń ze źródeł internetowych polega na zbudowaniu reprezentacji źródła oraz charakterystyk ekstrahowanego modelu służących wyznaczeniu zależności między wartościami zmiennych niezależnych modelu 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. Wykorzystując ekstrakcję informacji, nie tyle sięgamy do bazy danych udostępnionej przez stronę internetową, ile staramy się poznać algorytm (parametry i postać funkcji) wyliczania wielkości składki na podstawie zebranych danych. Zatem w odróżnieniu od ekstrakcji informacji ze źródeł głębokiego internetu, w zaproponowanym podejściu zajmujemy się ekstrakcją wiedzy. W prezentowanej części dokonujemy szczegółowego omówienia metody odtwarzania modelu, wyzwań z nią związanych oraz przedstawiamy narzędzie wspierające jej użycie. Ponadto pokazujemy uzyskane rezultaty, a także omawiamy zakres zastosowania.

Wprowadzenie

Druga część cyklu artykułów[1] wprowadziła istotne fragmenty proponowanej przez nas ontologii ubezpieczeniowej. Ponadto przedstawiono propozycje sposobów modelowania wiedzy o ryzyku jako szczególnym pojęciu zarówno dla teorii, jak też praktyki dziedziny ubezpieczeń.

1. W. Abramowicz, P. Stolarski, K. Węcel, *Ontologie jako narzędzie budowy modeli w ubezpieczeniowych systemach informacyjnych – modelowanie ryzyka oraz produktów cz. 2*, „Wiadomości Ubezpieczeniowe", nr 01/2011,

W niniejszej pracy w całości skupiamy się na zadaniu pozyskiwania modeli wyceny produktów ubezpieczeniowych ze źródeł internetowych. W związku z tym w części 1 prezentujemy wiadomości wstępne oraz wprowadzenie do problematyki. W części 2 dokonujemy przeglądu prac i rezultatów w związku z zagadnieniami wykazującymi pewne podobieństwo do przypadku będącego przedmiotem zainteresowania. W części 3 szczegółowo przedstawiamy prototyp metody, zastosowane ramy teoretyczne oraz ich praktyczne implementacje. W końcu w części 4, na podstawie otrzymanych do tej pory efektów, prezentujemy możliwe wykorzystanie ontologii ubezpieczeń oraz semantycznej informacji do niestandardowego przetwarzania wiedzy ubezpieczeniowej.

Przez wiedzę ubezpieczeniową rozumieć należy każdy zasób wiedzy bezpośrednio dotyczący rynku lub produktu ubezpieczeniowego. Szczególnym przypadkiem takiej wiedzy są modele wyceny produktów ubezpieczeniowych.

Odkrywaniu wiedzy ubezpieczeniowej ze źródeł internetowych towarzyszy szereg zagadnień, takich jak: ograniczenia zasobów, jakość pozyskanej wiedzy, a także jej aktualność. Metoda, której technologiczne aspekty opisane są poniżej, ma szereg potencjalnych zastosowań.

1. Ekstrakcja modeli wyceny ubezpieczeń ze źródeł internetowych – przedstawienie problemu

1.1. Zdefiniowanie problemu

Porównując proponowaną ekstrakcję modelu wyceny ze źródła internetowego z ekstrakcją informacji ze źródeł tzw. głębokiego internetu zauważyć można szereg pozornych podobieństw. Ze względu na różnorodność dziedzin skupimy się na różnicach między tymi zagadnieniami. Wskazać należy tutaj, iż:

- W odróżnieniu od ekstrakcji informacji w zaproponowanym podejściu ekstrahujemy wiedzę. Źródłem, do którego się odwołujemy, nie jest baza danych dostępna poprzez interfejs internetowy[2], lecz algorytm implementujący model wyceny.
- Po drugie, ilość uzyskiwanych poprzez zmianę kryteriów kalkulacji[3] wyników jednostkowych w przypadku ekstrakcji modeli wyceny może być zdecydowanie liczniejsza; ale też nie wszystkie wyniki jednostkowe przyczyniają się do podniesienia jakości ekstrahowanego modelu[4].
- W zaproponowanym podejściu ekstrakcja informacji jest częścią składową procesu ekstrakcji wiedzy.

Zadanie ekstrakcji modelu wyceny ubezpieczenia sprowadza się do odtworzenia jak najdokładniejszego[5] przybliżenia algorytmu wyceny składki. Uwzględnione są szczegółowe parametry polisy reprezentowane w postaci wektorów $\mathbf{X}_i = [x_{i1}, x_{i2}, \ldots, x_{im}]$ i odpowiadających im wycen s_i,

s. 117–137.

2. Najczęściej jest to pewien rodzaj formularza.
3. Ich odpowiednikiem w przypadku zadania ekstrakcji informacji z głębokiego internetu są kryteria wyszukiwania w bazie danych.
4. W przypadku istnienia zależności funkcyjnych na przedziałach kryteriów wyceny między kryteriami a wynikami jednostkowymi. Pojawia się jednak tutaj problem określenia, jaka informacja jest pomijana.
5. Przyjętą miarą dokładności jest MSE (Mean Standard Error), przy czym zakłada się, że błąd ten powinien przyjąć wartość mniejszą niż pewna określona wartość progowa.

$0<i\leq n$. Wyceny te są pozyskiwane w wyniku symulacji poruszania się (nawigacji) użytkownika po stronie internetowej ubezpieczyciela i wprowadzania zadanych parametrów polisy $x_{\bullet_1}, \ldots, x_{\bullet_m}$. Wartości parametrów wybierane są spośród dostępnych w formularzu internetowym opcji, określanych jako dziedzina parametru $x_{\bullet_i} \in D_j$, $0<j\leq m$. 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 poprzez 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. Na przykład, większa liczba zapytań zwiększa również ryzyko zablokowania dostępu do serwisu (działania antykonkurencyjne).

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

W przypadku opisywanego problemu, zarówno 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_i odtwarzanego modelu oraz ich dziedzin D_i,
- istnienie zdefiniowanej analitycznej postaci funkcji f[6].

W dalszej części artykułu pokazujemy, że możliwe jest opracowanie metody pozwalającej na wyekstrahowanie modelu ze zróżnicowanych źródeł internetowych[7] przy uwzględnieniu pewnych założeń minimalnych[8]. 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.

1.2. Identyfikacja internetowych źródeł wiedzy ubezpieczeniowej

Badanie przeprowadzone przez nas w lipcu 2011 roku pokazuje, że spośród 61 firm ubezpieczeniowych z siedzibą w Polsce oraz 17 oddziałów przedsiębiorstw zagranicznych wszystkie mają własne

6. Jest to silna wersja tego założenia. Słaba wersja założenia dopuszcza możliwość definiowania funkcji f za pomocą tablicy wartości.
7. Np. witryna internetowa, usługa sieciowa.
8. Założenia, o których tutaj mowa, dotyczą przede wszystkim technologicznych aspektów związanych z ekstrakcją informacji z rozpatrywanego źródła, takich jak: kwestie możliwości uzyskania dostępu, intensywności dostępu do źródeł, stabilności otrzymywanych informacji, niewystępowania przeszkód w komunikacji, etc.

witryny WWW. Nie są to wyniki zaskakujące w kontekście danych dotyczących wzrostu wolumenu sprzedaży produktów i usług przez internet na świecie[9, 10]. Dane te wyraźnie wskazują na rosnącą funkcję, zarówno marketingową, jak i sprzedażową, kanału internetowego. Jest to znaczny postęp w stosunku do roku 2006, gdy przeprowadzono podobne badanie. Wtedy 16,22 proc. podmiotów działu I oraz 9,09 proc. podmiotów działu II nie miało w ogóle witryny www[11].

Wykorzystanie internetu rośnie również w obszarze dystrybucji. Z naszych wyliczeń wynika, że w połowie 2011 roku w dziale II ubezpieczeń 57,69 proc. podmiotów z siedzibą w Polsce wykorzystywało kanał internetowy do bezpośredniej sprzedaży przynajmniej jednego produktu (w przeważającej częściej były to podmioty zorganizowane w formie spółek akcyjnych). Dla porównania w roku 2006 było to zaledwie 13,51 proc. podmiotów prowadzących działalność ubezpieczeniową w kraju[12].

Kolejnym przejawem wykorzystania internetu jest możliwość zgłoszenia szkód elektronicznie. Według naszych badań w ramach swojego modelu biznesowego 34,62 proc. podmiotów wspiera takie podejście. Przy czym, co można uznać za intrygujące, lista podmiotów umożliwiających zgłoszenie szkód poprzez internet nie pokrywa się z listą sprzedawców ubezpieczeń online.

Należy podkreślić, że rynek sprzedaży ubezpieczeń poprzez kanał internetowy nie sprowadza się wyłącznie do bezpośredniego oferowania ich przez ubezpieczycieli. W internecie obecni są także brokerzy, agenci, a także przedstawiciele nurtu *bancassurance*. Na popularności powoli zyskują także portale porównujące oferty.

Tabela 1. pokazuje, jaki procent produktów w poszczególnych grupach ubezpieczeń oferowany jest do bezpośredniej sprzedaży w internecie.

Tabela 1. Stopień przeniesienia grup ubezpieczeń do oferty w kanale internetowym[13]

OC	AC	NNW	DOM	TURYSTYKA
60,00%	60,00%	40,00%	33,33%	40,00%

Źródło: opracowanie własne.

1.3. Przykłady źródeł danych dla ekstrakcji modeli wyceny ubezpieczeń

W niniejszej sekcji krótko przedstawiamy opis źródeł z wykorzystaniem formalizmów utworzonych na potrzeby metody stanowiącej przedmiot artykułu. Dokładniejsze omówienie zademonstrowanych formalizmów zawarte jest w sekcji 3.1.

Nawigacja po pierwszym źródle – benefia24.pl – zdefiniowana może być przez graf pokazany na rysunku 2. Z kolei kod XML przedstawiony na rysunku 1 przedstawia fragment definicji dynamicznego drzewa – struktury przeznaczonej do opisu funkcjonowania i dopuszczalnych wartości parametrów w dynamicznym formularzu występującym na jednej z podstron w procesie pozyskiwania danych.

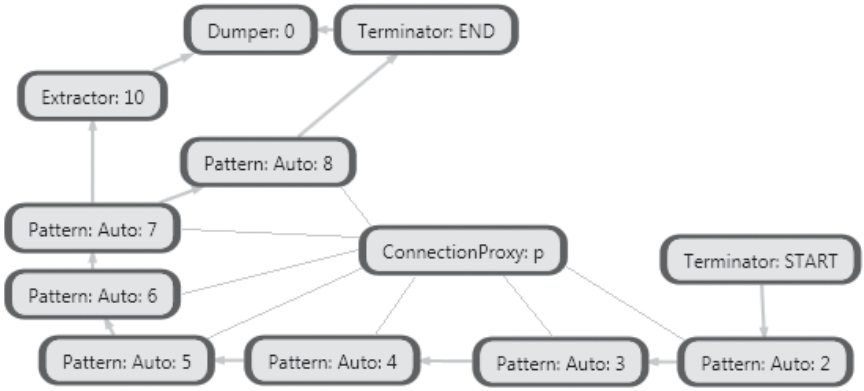
9. Patrz np. „US Online Insurance Forecast, 2010 To 2015", Forrester Research, Inc. 2011.
10. http://www.bankier.pl/wiadomosc/Co-kilka-sekund-ktos-dzwoni-lub-klika-po-polise-2427449.html (12.12.2011).
11. M. Kaczała, *Internet jako instrument dystrybucji ubezpieczeniowej,* praca doktorska, UEP 2006.
12. Wzrost w zakresie przystosowania modelu sprzedaży w ciągu 5 lat wyniósł przeszło 320 proc.
13. Dane procentowe odnoszą się do populacji firm ubezpieczeniowych w formie spółek akcyjnych, prowadzących sprzedaż online.

Rysunek 1. Fragment kodu XML

```
<SerializableKeyValuePairOfStringObject>
  <Key>emotoplaceholderpojazdviewcarselectorddlpojemnosc</Key>
  <Value xsi:type="ConditionalTextualIterableProperty">
    <ConditionalValues>
      <NodeOfConditionalValue><Data><Condition>
      <Value xsi:type="xsd:string">5144</Value>
      <PropertyName>&lt;emotoplaceholderpojazdviewcarselectorddltyp&gt;
      </PropertyName>
      <Relation>=</Relation></Condition></Data>
        <Children><NodeOfConditionalValue><Data><Value>2245</Value>
        </Data></NodeOfConditionalValue></Children>
        </NodeOfConditionalValue>
      <NodeOfConditionalValue><Data><Condition>
      <Value xsi:type="xsd:string">5145</Value>
      <PropertyName>&lt;emotoplaceholderpojazdviewcarselectorddltyp&gt;
      </PropertyName>
      <Relation>=</Relation></Condition></Data>
        <Children><NodeOfConditionalValue><Data><Value>2997</Value>
        </Data></NodeOfConditionalValue></Children>
        </NodeOfConditionalValue>
```

Źródło: opracowanie własne.

Rysunek 2. Graf opisujący nawigację po przykładowym źródle wiedzy ubezpieczeniowej



Źródło: opracowanie własne.

Rysunek 3. Fragment kodu XML

```
<Properties>
  <SerializableKeyValuePairOfStringObject>
    <Key>ctl00$contentmain$ctl00$txtboatvalue</Key>
    <Value xsi:type="NumericIterableProperty">
      <Id>ctl00$contentmain$ctl00$txtboatvalue</Id>
      <Minimum>10000</Minimum>
      <Step>10000</Step>
      <Maximum>50000</Maximum>
      <Current>30000</Current>
    </Value>
  </SerializableKeyValuePairOfStringObject>
  <SerializableKeyValuePairOfStringObject>
    <Key>ctl00$contentmain$ctl00$ddlmooring</Key>
    <Value xsi:type="TextualIterableProperty">
      <Id>ctl00$contentmain$ctl00$ddlmooring</Id>
      <TextualValues>
        <string>MARINA</string>
        <string>HARBOUR</string>
        <string>ASHORE</string>
        <string>LOCKED</string>
        <string>BANKSIDE</string>
        <string>OTHER</string>
      </TextualValues>
      <Current>ASHORE</Current>
    </Value>
```

Źródło: opracowanie własne.

Drugi przykład źródła to http:/insurance-4-boats.co.uk. Źródło to ma prostą strukturę nawigacyjną, z prostym interfejsem. Kod XML przedstawiony na rysunku 3 odpowiedzialny jest za deklarację części wykrytych parametrów modelu. Parametry w odróżnieniu od pokazanego w poprzednim przykładzie dynamicznego drzewa zdefiniowane są globalnie dla całego źródła.

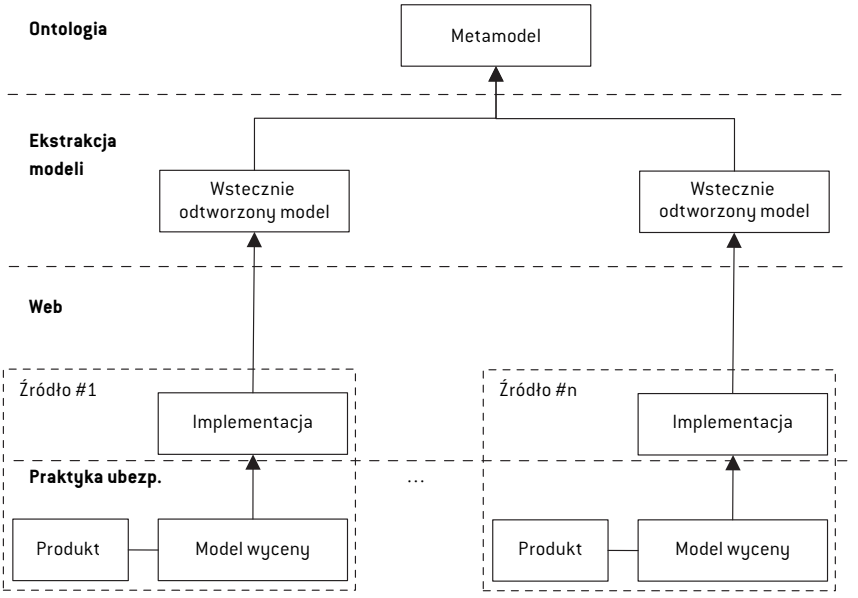
1.4. Przypadki szczególne

Przedstawiony w sekcji 1.1 problem ogólny ma wiele przypadków szczególnych oraz problemów powiązanych:

- budowa metamodeli reprezentujących wycenę składek dla grupy zbliżonych produktów ubezpieczeniowych,
- porównanie modeli wyceny zbliżonych produktów ubezpieczeniowych,
- uwzględnienie zmienności danych, parametrów i modeli w czasie,
- określenie znaczenia oraz istotności parametrów w modelu.

Problem porównania modeli wyceny wymaga operowania na wyższym poziomie abstrakcji niż poziom technologii ekstrakcji informacji. Naturalnym sposobem agregacji aparatu pojęciowego wydaje się ontologia dziedzinowa. Poprzez metamodel rozumiemy tutaj pewien wzorcowy lub uogólniony model wyceny, zawierający np. rozszerzoną liczbę parametrów lub łączący zależności grup zbliżonych parametrów wraz z ich wpływem na cenę składki[14]. Hierarchia modeli wyceny aż do poziomu metamodelu zaprezentowana jest na rysunku 4. Model taki może mieć znaczenie referencyjne jako osobny zasób wiedzy.

Rysunek 4. Poziomy operowania na modelach wyceny



Źródło: opracowanie własne.

14. G. Dionne, C. Vanasse, *A generalization of actuarial automobile insurance rating models. The negative binomial distribution with a regression component*, „Actuarial Bulletin" nr 19/1989, s. 199–212.

2. Prace powiązane

2.1. Modelowanie wyceny ryzyka w produktach ubezpieczeniowych

Metody tworzenia produktów ubezpieczeniowych różnią się w istotny sposób między działami ubezpieczeń, chociaż ostateczny cel konstrukcji ubezpieczeń w zasadzie jest niezależny od rozróżnienia między ubezpieczeniami związanymi z życiem lub zdrowiem oraz ubezpieczeniami majątkowymi. Odmienne metody są także charakterystyczne wewnątrz danego działu ubezpieczeń. Wynika to m.in. z dostępności danych, różnic w charakterze ryzyk, uregulowań prawnych dotyczących produktów, zobowiązań umownych lub założeń modelujących.

Wspólny jest niewątpliwie ogólny schemat przygotowania wyceny: zebranie wstępnych danych; kalkulacja składki netto; uwzględnienie doliczeń i odliczeń; kalkulacja składki brutto; obliczenie rezerw. Istotnym aspektem technik aktuarialnych jest także ciągły ich rozwój m.in. poprzez włączanie nowych technik (symulacje[15], data mining[16]).

Istotną informacją przy tworzeniu nowego produktu lub uaktualnianiu istniejącego jest dokumentacja ratingu, niezbędna dla procesu obliczania składki dla danego ryzyka. Dokumentacja taka sprowadza się do: wyznaczenia reguł, tabel klasyfikacyjnych i opłat, algorytmu ratingowego oraz ewentualnych uwag do procesu underwritingu.

Stosowane metody w przypadku ubezpieczeń na życie są zróżnicowane przede wszystkim ze względu na szczegółowy typ ubezpieczenia oraz uregulowań umownych. Charakterystycznymi cechami modeli wyceny ubezpieczeń w tym dziale są: intensywne wykorzystanie matematyki finansowej dla uwzględnienia efektu zmiany wartości pieniądza w czasie oraz wykorzystanie danych zagregowanych w ramach tzw. tablic trwania życia jako bazy statystycznej. Dobre omówienie tak tworzonych modeli można znaleźć w wielu opracowaniach naukowych[17, 18].

W przypadku ubezpieczeń majątkowych obliczanie i konstrukcja płatności składek wykonywana jest za pomocą zróżnicowanych metod. Do najważniejszych z nich zaliczyć należy:

- metodę tradycyjną klasyfikacji ryzyk (jednej zmiennej)[19]: czysta składka; współczynnik strat; skorygowana czysta składka; modele multiplikatywne, tj. minimum Chi^2 (Bailey-Simon), marża całkowita, maksymalne prawdopodobieństwo[20],
- modele klasyfikacji wielu zmiennych: uogólniony model liniowy[21] (GLM); techniki wspomagające: data-mining (sieci neuronowe, MARS, CART, analiza skupień, analiza czynnikowa),
- podejścia specjalne, obejmujące: analizę terytorialną; czynniki zwiększonych limitów; odliczenia[22]; ubezpieczenie na wartości/ko-ubezpieczenie,

15. R. Salam, *Estimating the Cost of Commercial Airlines-Catastrophes A Stochastic Simulation Approach*, CAS Forum, Winter 2003.
16. Metody te stanowią również temat bezpośredniego zainteresowania ze względu na omawianą w artykule metodę, w związku z tym skrótowe omówienie pojawi się w kolejnych sekcjach.
17. B. Błaszczyszyn, T. Rolski, *Podstawy matematyki ubezpieczeń na życie*, WNT, Warszawa 2004.
18. H.U. Gerber, *Life Insurance Mathematics*, Springer-Verlag 1997 (3 wyd.).
19. G.Werner, C. Modlin, *Basic Ratemaking, 4th ed.*, Casualty Actuarial Society, 2010.
20. A. Björn, *Comparison of Some Methods to Fit a Multiplicative Tariff Structure to Observed Risk Data*, „ASTIN Bulletin International Actuarial Association", 1968, vol. 16, nr 1, Belgia, s. 63–68.
21. D. Anderson, et. al., *A Practitioner's Guide to Generalized Linear Models*, „Casualty Actuarial Society Study Note", May 2005.
22. B.Z. Brown, M.C. Schmitz, *Study Note on Deductibles*, „CAS Study Note", July 2006.

- podejście indywidualne[23]: plany dużych odliczeń; złożony rating oceny strat; plany ratingu retrospektywnego.

Zadaniem komplementarnym do części z powyższych metod jest dobór procedury wiarygodności[24].Najczęściej stosowanymi metodami są tutaj: wiarygodność wg najmniejszych kwadratów (Bühlmann); klasycznie – ograniczona zmienność; analiza bayesowska.

Szczególną grupą ubezpieczeń są polisy ubezpieczające od zdarzenia, a nie wystawiane tylko na okres obowiązywania umowy. Ze względu na opóźnienia związane z czasem zgłoszenia oraz procedur poprzedzających wypłaty, odszkodowania te wymagają szczególnego potraktowania, w tym m.in. zastosowania odpowiednich technik agregacyjnych dla danych.

2.2. Ekstrakcja informacji z internetu

Informacja w sieci Web występuje w postaci ustrukturyzowanej lub nieustrukturyzowanej. Informacja ustrukturyzowana w znaczącym stopniu charakteryzuje się regularnością wykorzystanej formy, co umożliwia jej automatyczne przetwarzanie. Informacja częściowo ustrukturyzowana lub nieustrukturyzowana udostępniana jest w postaci wymagającej interpretacji przez ludzi. Automatyczne przetwarzanie takiej informacji jest utrudnione lub wręcz niemożliwe.

Jeśli chodzi o podejścia do strukturyzacji informacji, to wyodrębnić można dwa zasadnicze nurty: oddolny i odgórny. W pierwszym przypadku twórcy treści (stron) są odpowiedzialni za oznaczenie tekstu tak, żeby był łatwo przetwarzalny w sposób automatyczny[25]. Nurt odgórny nie zakłada zmiany sposobu publikowania informacji w sieci. W zamian proponuje on zastosowanie algorytmów tzw. eksploracji Web[26] do wykrywania i ekstrahowania informacji ze źródeł sieciowych. Należy stwierdzić, że pierwszy nurt jest względnie rzadko spotykany, ponieważ: twórcy traktują człowieka jako podstawowego odbiorcę treści. Duża część treści w sieci ma charakter historyczny. Wiele firm (m.in. w sektorze handlu elektronicznego) dąży do utrudnienia dostępu do informacji odbiorcom niebędącym bezpośrednimi klientami.

Serwisy zasilane danymi to źródła internetowe przedstawiające informacje najczęściej w postaci półstrukturalnej. Stanowią one bardzo szeroką kategorię źródeł spotykanych w sieci www. Bardziej szczegółowo ze względu na technologię wyodrębnić wśród nich można grupy:

- proste serwisy zasilane danymi – korzystające z tabel, list i znaczników języka HTML do prezentacji danych, w postaci statycznych dokumentów o stosunkowo prostej i powtarzalnej strukturze,
- serwisy o zaawansowanym GUI[27] – wykorzystujące technologie[28] dynamicznych interfejsów użytkownika, umożliwiające wygodną dla odbiorców prezentację informacji,
- źródła głębokiego internetu – dające dostęp do danych tylko po wypełnieniu formularzy HTML, często zawierających tzw. pola otwarte (np. wymagane pola tekstowe),

23. G.Werner, C. Modlin, op. cit.
24. H.C. Mahler, C.G. Dean, *Credibility, Chapter 8 in Foundations of Casualty Actuarial Science 4th ed.*, Arlington, VA: Casualty Actuarial Society, 2001.
25. Wykorzystane mogą być w tym celu różne formalizmy, takie jak: XML, XSLT, RSS, RDF, RDFa, Microformats, OWL, JSON, DublinCore i inne techniki wspomagających strukturalizację informacji.
26. Ang. *Web mining*.
27. Graphical User Interface – graficzny interfejs użytkownika.
28. Np. AJAX, Flash.

- serwisy spersonalizowane – uzależniające wyświetlane treści od ustawień i innych charakterystyk zalogowanego użytkownika,
- aplikacje sieci Web – charakteryzujące się stanowością, tj. przechowujące pewne informacje pomiędzy kolejnymi połączeniami z daną witryną (w ramach jednej lub wielu sesji),
- inne, m.in. serwisy adaptatywne, serwisy wykorzystujące filtrowanie grupowe, sieci społecznościowe, uogólnione serwisy zasilane danymi.

Analizowane w artykule źródła mają najczęściej cechy kilku grup. Najczęściej rozpatrywanym w literaturze, i najistotniejszym z punktu widzenia niniejszego artykułu, jest problem pozyskiwania (ekstrakcji) informacji właśnie z serwisów zasilanych danymi oraz zagadnienia powiązane: monitorowanie[29] i integracja informacji z takich źródeł.

W tradycyjnych podejściach[30][31]ekstrakcja dokonywana jest poprzez transformację półustrukturyzowanych dokumentów w języku HTML do formy w pełni strukturalnej, np. relacyjnej.

Pół- lub całkowicie automatyczne ekstrahowanie informacji ze źródeł sieci Web wymaga wewnętrznego mechanizmu reprezentacji takiego źródła. Osłoną[32]określa się komponent programowy pozwalający na uogólnione podejście do takiej reprezentacji. Systemy ekstrakcji ze źródeł internetowych można podzielić ze względu na charakterystyczne cechy i technologię używanych osłon następująco:

- prymitywne osłony w postaci ręcznie tworzonych reguł[33] – TSIMMIS, Ariadne czy Web-OQL[34, 35, 36],
- silniki opakowujące działające na danych dostarczonych przez użytkownika poprzez specjalny interfejs użytkownika – NoDoSE[37], W4F[38],
- osłony tworzone dzięki technikom uczenia maszynowego z przykładami – WIEN[39],
- osłony budowane za pomocą technik uczenia nienadzorowanego – Exalg[40].

W przypadku skomplikowanych źródeł sieci Web reprezentacja źródła wymaga bardziej wyrafinowanego podejścia. Sytuacja taka dotyczy przede wszystkim źródeł głębokiego internetu oraz serwisów o zaawansowanym GUI. Źródła głębokiego internetu generują dodatkowo problem

29. Zagadnienia monitorowania źródeł i integracji informacji zostały celowo pominięte w niniejszym opracowaniu.
30. L. Eikvil. *Information extraction from world wide web – a survey*, raport instytutowy, 1999.
31. A. McCallum , W.W. Cohen, *Information extraction from the world wide web, tutorial*, 2002.
32. Ang. *wrapper*, tłumaczenie za K. Subietą.
33. Przede wszystkim rozwiązania pionierskie lub wysoce specjalistyczne.
34. S. Chawathe, Y. Papakonstantinou, J.D. Ullman, H. Garcia-Molina, K. Ireland, J. Hammer, J. Widom, *The TSIMMIS project: Integration of heterogeneous information sources*, 10th Meeting of the Information Processing Society of Japan, 1994, s. 7–18.
35. C.A. Knoblock, S. Minton, J.L. Ambite, N. Ashish, I. Muslea, A.G. Philpot, S. Tejada, *The ariadne approach to web-based information integration*, „International Journal of Cooperative Information Systems", nr 10(1)/2001, s. 145–169.
36. G.O. Arocena, A.O. Mendelzon, *Weboql: Restructuring documents, databases, and webs*, 14th International Conference on Data Engineering, 1998.
37. 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, 1998, s. 283–294.
38. F. Azavant, A. Sahuguet, *Bulding light-weight wrappers for legacy web data sources using w4f*, 25th International Conference on Very Large Data Bases, 1999.
39. N. Kushmerick, *Wrapper induction for information extraction*, praca doktorska, University of Washington, 1997.
40. A. Arasu, H. Garcia-Molina, *Extracting structured data from web pages*, 2003 ACM SIGMOD International Conference on Management of Data, 2003, s. 337–348.

nawigacji poprzez formularze[41,42]. Z kolei nawigowanie po źródłach z zaawansowanym GUI wymaga m.in. pokonania wyzwania, jakim są dynamicznie zmieniane treści[43].

2.3. Ekstrakcja wiedzy z wykorzystaniem metod *data mining*

Metody eksploracji danych (data mining) służą do odkrywania ukrytych wzorców, powiązań i trendów przez przeszukiwanie ogromnych ilości danych z wykorzystaniem odpowiednich metod statystycznych. Celem ich stosowania jest identyfikowanie nowych, potencjalnie użytecznych oraz zrozumiałych wzorców w zgromadzonych danych. Powiązanym pojęciem jest odkrywanie wiedzy (*knowledge discovery*), które dotyczy pozyskiwania wiedzy odnoszącej się do organizacji na podstawie danych zgromadzonych w zasobach danych w celu wykorzystania jej do podejmowania decyzji biznesowych. Dostarcza szeregu technik, które pozwalają na ekstrakcję wiedzy.

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*[44]. Pierwszy z nich to dokonywanie wyborów, czyli klasyfikacje, dla których szczególne zastosowanie mają drzewa decyzyjne[45] i analiza skupień[46]. Drugi problem to tworzenie prognoz, czyli przybliżanie nieznanej przyszłej wartości. 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 klasyczna analiza regresji. Z punktu widzenia problemu naukowego niniejszego artykułu zagadnienie tworzenia takich modeli predykcyjnych jest szczególnie istotne. Trzeci problem to odkrywanie relacji w danych, a czwarty to usprawnianie procesów.

Zarówno badania nad sieciami neuronowymi, jak i algorytmami ewolucyjnymi, zostały zainspirowane obserwacjami biologicznymi. Pierwsze matematyczne modele neuronów powstały w latach czterdziestych ubiegłego wieku.[47] Modele te od tego czasu są systematycznie ulepszane z uwzględnieniem specyficznych cech funkcjonowania komórek nerwowych. Dla budowy sieci neuronowych poza modelem takiej komórki szczególnie ważny jest również sposób powiązania

41. D. Shestakov, S.S. Bhowmick, Lim E., *Deque: Querying the deep web*, „Data Knowl. Eng.", nr 52(3)/2005, s. 273–311.
42. T. Kaczmarek, *Integracja danych z głębokiego Internetu dla potrzeb analizy otoczenia przedsiębiorstwa*, praca doktorska, Akademia Ekonomiczna w Poznaniu, 2006.
43. 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, 2007, s. 18–25.
44. M.J.A. Berry, G.S. Linoff, *Mastering Data Mining.The Art and Science of Customer Relationship Management*, Wiley Computer Publishing, New York 2000, s. 494.
45. J.R. Quinlan, *Induction of Decision Trees*, „Machine Learning", nr 1/1986, Morgan Kaufmann, s. 81–106.
46. M. Bramer, *Principles of Data Mining*, Springer, London 2007.
47. W.S. McCulloch, W.H. Pitts, *A logical calculus of the ideas immanent in nervous activity*, „Bulletain of Mathematical Biophysics", nr 5/1943, s. 115–133.

pojedynczych neuronów w sieć. Wyróżniono zatem szereg konfiguracji charakterystycznych dla poszczególnych typów sieci[48]. Kolejnym istotnym zagadnieniem jest sposób uczenia sieci[49].

Algorytmy ewolucyjne podzielić można na algorytmy genetyczne[50], 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[51]. Koncepcja programowania genetycznego rozwinięta została przez J. Kozę[52].

Idea programowania genetycznego może być zrealizowana w zróżnicowany sposób. W ujęciu klasycznym generowane programy zapisane były w języku LISP, a więc miały strukturę drzew, gdzie węzły zawierały funkcję, a liście jej argumenty (możliwe jest zatem składanie funkcji). Obecnie częstym zabiegiem jest wykorzystanie dla reprezentacji programów pseudokodu maszynowego (metoda AIM[53]). Kod taki może być później tłumaczony na języki wysokiego poziomu. W programowaniu genetycznym kolejne programy przybliżające funkcje generowane są poprzez zastosowanie operatorów ewolucyjnych na populacji początkowo losowo stworzonych programów. Operatorami ewolucyjnymi są klasycznie mutacja oraz krzyżowanie, przy czym intensywność ich stosowania określona jest odpowiednimi parametrami. Istotną różnicą w stosunku do innych algorytmów ewolucyjnych jest definiowanie podstawowych jednostek podlegających ewolucji, tj. genów i chromosomów. W przypadku programowania genetycznego genami są elementy budujące funkcję (operatory arytmetyczne, trygonometryczne, funkcje warunkowe, etc. oraz ich argumenty). Chromosomami są zakodowane funkcje, przy czym zarówno stopień skomplikowania elementów budujących, jak i długość chromosomów zależy od szeregu parametrów oraz konkretnych implementacji algorytmu.

Działanie algorytmu jest następujące. Najpierw tworzona jest początkowa populacja programów P. Poprzez zastosowanie operatorów mutacji i krzyżowania dla każdego osobnika z P powstaje nowa populacja P'. Obie populacje podlegają łącznie ocenie za pomocą funkcji błędu, w celu wyboru najlepszych osobników (programów), tj. pozwalających na najlepsze przybliżenie nieznanej wartości. Osobniki znajdujące się najwyżej w rankingu przechodzą do nowej populacji P, a algorytm jest powtarzany, przy założeniu, że nie zostało znalezione satysfakcjonujące rozwiązanie.

3. Opis metody ekstrakcji wiedzy ze źródeł internetowych – wybrane zagadnienia

3.1. Źródło internetowe – reprezentacja

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

48. G.F. Miller, P.M. Todd, S.U. Hagde, *Designing neural networks using genetic algorithms*, Proc. of the 3rd Int. Conf. on Genetic Algorithms and Their Applications, [red.] J.D. Schaffer, Morgan Kaufmann, San Mateo, USA 1989, s. 379–384.
49. D. Montatna, L. Davis, *Training feedforward neural networks using genetic algorithms*, Proc. of the 11th Int. Conf. on ArtificialIntelligence, Morgan Kaufmann, USA 1989, s. 762–767.
50. J. Arabas, *Wykłady z algorytmów ewolucyjnych*, WNT, Warszawa 2001.
51. A.S. Fraser, *Simulation of genetic systems by automatic digital computers*, „J. Biol. Sci.", nr 10/1957, s. 484–499.
52. J.R. Koza, *Genetic Programming II: Automatic Discovery of Reusable Programs*, The MIT Press, 1994.
53. S. Mukkamala, A.H. Sung, B. Ribeiro , A.S. Vieira, J.C. Neves, *Model Selection and Feature Ranking for Financial Distress Classification*, Proc. of 8th Int. Conf. on Enterprise Information Systems (ICEIS), 2006.

internetowe (więcej wcześniej w sekcji 2.2). Standardowym rozwiązaniem jest wielostronicowy formularz z nawigacją. Jednocześnie nie są rzadkością rozwiązania polegające na wykorzystaniu mechanizmów asynchronicznego transferu informacji (AJAX) oraz intensywnego wykorzystania języka JavaScript. Często można się spotkać nie tyle z przesyłaniem dodatkowych danych za pomocą technologii AJAX, ile z tworzeniem w sposób dynamiczny całych sekcji formularzy. Takie rozwiązanie z pewnością odpowiada potrzebom funkcjonalnym związanym z wymogiem uzyskania szczegółowej informacji od użytkownika.

Z drugiej strony część tego typu witryn posiada 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 liczby połączeń z określonych węzłów sieci. Do grupy drugiej zaliczyć można m.in. mechanizmy „Captcha"[54], konieczność rejestracji lub identyfikację za pomocą danych personalnych. Zaimplementowany przez nas prototyp badawczy pozwolił na częściowe ominięcie powyżej opisanych mechanizmów, dzięki czemu możliwe było zebranie danych do badań.

W celu zbudowania skutecznego narzędzia ekstrakcji danych z portalu ubezpieczeniowego konieczne było stworzenie zaawansowanego modelu takiego portalu w celu automatycznej po nim nawigacji.

Model źródła portalu sprzedaży ubezpieczeń zastosowany dotychczas w naszym prototypie badawczym ma dwa wymiary. Pierwszy wymiar dotyczy opisu źródła w zakresie nawigacji między 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.

Nawigacja między lokalizacjami reprezentowana jest za pomocą kolorowanego, skierowanego grafu, w którym poszczególne typy wierzchołków odpowiadają artefaktom związanym z nawigacją po źródle, np. podstrona, miernik czasu, serwer proxy, warunek logiczny.

Dla opisu drugiego wymiaru posłużono się specyficzną strukturą, którą określić można jako „drzewo warunkowe". Jest to drzewo o krawędziach oznaczonych 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.

3.2. Ekstrakcja danych

W opisie zadania ekstrakcji modelu wyceny zawartym w sekcji 1.1 stwierdzono, że ważną cechą metody jest redukcja liczby zapytań do źródła (serwera), w ramach których otrzymywane są wartości kolejnych składek $f(\mathbf{X}_i)$. W zakresie strategii optymalizacji ekstrakcji wycen ze źródła wyróżnić można kilka podejść:

a) brak optymalizacji – podejście naiwne,
b) naiwne odkrywanie liniowości przy założeniu *ceteris paribus*,
c) naiwne odkrywanie braku wpływu parametru polisy,
d) posiadanie dodatkowej wiedzy o parametrze polisy,

54. 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.

e) posiadanie dodatkowej wiedzy o modelu,
f) odkrywanie zależności funkcyjnych lub braku wpływu poprzez wnioskowanie statystyczne przy założeniu *ceteris paribus*,
g) odkrywanie zależności funkcyjnych lub braku wpływu poprzez wnioskowanie statystyczne z uwzględnieniem wpływu zmian innych zmiennych niezależnych modelu.

W omawianym prototypie metody zaimplementowano rozwiązania wymienione w punktach b)–f).

Szacunkowa liczba wszystkich możliwych wartości parametrów testowych modeli w przeprowadzonym badaniu wahała się w granicach od 144000 do $8{,}6*10^9$. Taka liczba zapytań wyklucza strategię braku optymalizacji (strategia a).

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

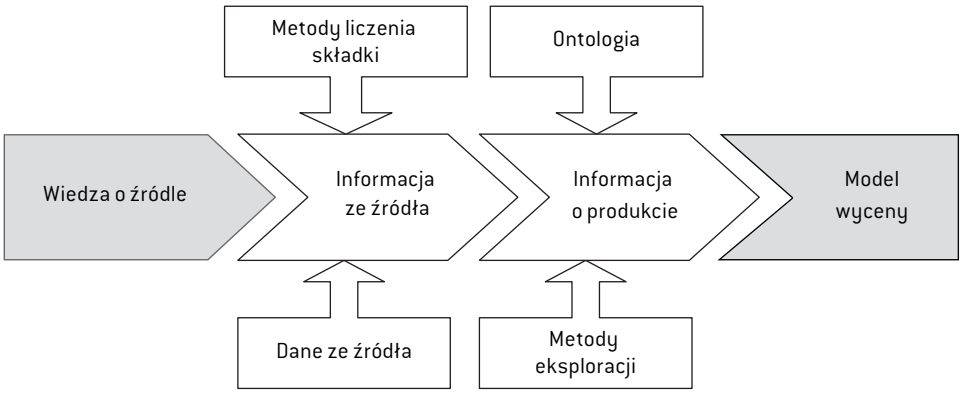
Strategie d) i e) w eksperymentalnym prototypie zostały zrealizowane poprzez wprowadzenie możliwości ręcznego przypisania pojęć z ontologii dziedzinowej do poszczególnych elementów modelu źródła internetowego. Użyto odpowiednio rozszerzoną ontologię w języku OWL zaprezentowaną w poprzedniej części artykułu[55]. Ta dodatkowa wiedza pozwala usprawnić działanie systemu ekstrahującego, np. poprzez przypisanie predefiniowanego zbioru wartości dla określonego parametru polisy.

Strategia f) 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.

3.3. Przekształcenie danych w wiedzę

Istotną cechą proponowanej metody jest ekstrakcja wiedzy o produkcie ubezpieczeniowym w opozycji do ekstrakcji informacji ze źródeł internetowych prezentowanej w sekcji 2.2. Zebranie danych ze źródła jest tylko jednym z etapów. Całość procesu została przedstawiona na Rysunku 5.

Rysunek 5. Obieg wiedzy w ramach prezentowanej metody ekstrakcji



Źródło: opracowanie własne.

55. W. Abramowicz, P. Stolarski, K. Węcel, op. cit., s. 117–137.

Dzięki zastosowaniu prototypu metody wraz z dedykowanymi jej narzędziami następuje łączenie wiedzy zaimplementowanej w ramach systemu ekstrahującego wraz z informacją pochodzącą ze źródła webowego. W rezultacie kolejnych przekształceń otrzymujemy na koniec procesu nową wiedzę w postaci modelu wyceny produktu ubezpieczeniowego. Proces taki dobrze odpowiada niektórym definicjom wiedzy[56].

Zagadnienie przekształcania danych w wiedzę za pomocą proponowanej metody związane jest z problemem modelowania źródła danych *sensu stricte*, a więc źródła, które generuje ekstrahowane później dane.

W przypadku zastosowań związanych z ekstrakcją informacji w roli źródła najczęściej występuje baza danych. Modelowanie takiego źródła sprowadza się do próby opisania samej bazy danych. Przykład można znaleźć w projekcie WISE-Integrator[57], gdzie próbowano dokonać integracji informacji przy wykorzystaniu analizy schematu interfejsu internetowej bazy danych.

W przypadku rozważanej metody problem ma bardziej ogólną postać – nie ma pewności, że istnieje baza danych, która generuje otrzymywane w wyniku wykonywanych zapytań rezultaty. Dużo bardziej prawdopodobne jest, że mamy do czynienia z modelem wyliczającym składki na bieżąco. Może on przyjmować postacie: drzew decyzyjnych, modeli analitycznych lub modelu mieszanego, który obejmuje zarówno bazę z danymi, jak i określony model analityczny. Takie rozróżnienie wskazuje, że rozwiązanie problemu ekstrakcji modeli wyceny wymaga bardziej ogólnego podejścia do modelowania źródła danych *sensu stricte* niż klasyczne podejście ekstrakcji informacji z głębokiego internetu.[58]

3.4. Ekstrakcja modeli wyceny produktu ubezpieczeniowego ze źródeł internetowych

Koncepcja procesu ekstrakcji modeli wyceny produktów ubezpieczeniowych ze źródeł internetowych została zaprezentowana w poprzedniej części artykułu[59]. W jej realizacji praktycznej wprowadzono także usprawnienia wynikające z ograniczeń zasobów oraz dostępności komponentów programowych wykorzystanych w rozwiązaniu. Etapy realizacji praktycznej wraz z ich cechami charakterystycznymi przedstawia Tabela 2.

Tabela 2. Proces ekstrakcji modeli wyceny produktu ubezpieczeniowego ze źródła internetowego

Faza	Etap	Automatyzacja	Wsparcie	Uwagi
Przygotowanie				
	Test źródła	Brak	–	
	Deklaracja kandydatów na parametry modelu	Pełna	Przez interfejs użytkownika	

56. Można się spotkać z definicją, iż wiedza to każdy zasób, który pozwala uzyskać nowe zasoby wiedzy poprzez złożenie posiadanych zasobów z nową informacją (A.M. Gadomski, *ENEA's e-paper*, http://erg4146.casaccia.enea.it/wwwerg26701/gad-dict.htm).
57. H. He, W. Meng, C. Yu, Z. Wu, *Wise-integrator: A system for extracting and integrating complex web search interfaces of the deep web*, Proceedings of 31st International Conference on Very Large Data Bases, 2005.
58. Dokładniej rzecz biorąc, uogólnienie polega na tym, że generator w postaci bazy danych jest tylko bardzo szczególnym przypadkiem źródła.
59. W. Abramowicz, P. Stolarski, K. Węcel, op. cit., s. 117–137.

	Budowa grafu nawigacji	Pełna	Przez interfejs użytkownika	
	Uszczegółowienie grafu nawigacji	Częściowa	–	Edycja XML
	Oznaczenie pojęciami sterującymi	Brak	–	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	–	

Źródło: opracowanie własne.

4. Wstępne wyniki i dyskusja nad zastosowaniami

4.1. Otrzymane wyniki

W ramach pilotażowego eksperymentu pozyskano i skutecznie wyekstrahowano dane pochodzące z pięciu modeli wyceny oraz trzech niezależnych źródeł. W tym celu stworzone zostało prototypowe oprogramowanie zbudowane z wykorzystaniem ogólnodostępnych, otwartych komponentów programistycznych[60, 61]. W części analitycznej wykorzystano moduł Enterprise Miner pakietu SAS.[62] Tabela 3 przedstawia ilościowe zestawienie wyekstrahowanych danych.

Tabela 3. Liczba rekordów pozyskanych z poszczególnych modeli w eksperymencie

Źródło	Model	Liczba rekordów
Źródło 1	Model 1	321
Źródło 2	Model 2	1943
Źródło 2	Model 3	1085
Źródło 3	Model 4	1300
Źródło 3	Model 5	1200

Źródło: opracowanie własne.

W obrębie źródła pierwszego zagregowano dane opisujące jeden model stawki bazowej ubezpieczenia na życie oferowanego przez jeden z czołowych banków ze Stanów Zjednoczonych. Ze względu na wstępny charakter badań zebrano dane z modelu uproszczonego, bez uwzględniania dużej liczby zmiennych binarnych opisujących stan zdrowia ubezpieczanego. Odtworzenie pełnego modelu byłoby znacznie trudniejsze.

W przypadku analizy drugiego źródła otrzymano wyniki dwóch konkurencyjnych modeli wyceny ubezpieczenia statków wodnych określonej klasy oferowanych przez czołowych ubezpieczycieli na rynku brytyjskim. Zebranie wyników było utrudnione, gdyż w źródle wprowadzono zabezpieczenia przed zbyt intensywnym wykorzystaniem źródła.

60. http://code.google.com/p/csexwb2/(20.11.2011).
61. http://owlapi.sourceforge.net/(20.11.2011).
62. http://www.sas.com/technologies/analytics/datamining/miner/(20.11.2011).

Źródło trzecie reprezentuje polski sektor ubezpieczeniowy. Wybrano dwa modele opisujące produkt ubezpieczenia OC pojazdu mechanicznego oraz ubezpieczenie nieruchomości przed skutkami takich ryzyk, jak: pożar, kradzież z włamaniem, rozbój, inne zdarzenia losowe, OC w życiu prywatnym. Również tutaj zaistniała potrzeba przezwyciężenia zabezpieczeń. Różnice w liczbie rekordów pobranych z poszczególnych źródeł odzwierciedlają skomplikowanie danego modelu ze względu na liczbę zmiennych. Nadmiarowe dane wykorzystano do zwiększenia zbiorów walidacyjnych.

Tabela 4. Charakterystyka danych zebranych w trakcie eksperymentu

Model	Liczba zmiennych	Liczba zmiennych istotnych	Odchylenie standardowe składek	Rozpiętość składek
Model 1	6	6	167,15$	1127,72$
Model 2	11	7	£94,12	£306,47
Model 3	11	7	£37,90	£273,98
Model 4	17	12	1444,24 zł	7557,05 zł
Model 5	17	14	314,04 zł	1466,88 zł

Źródło: opracowanie własne.

Tabela 5. Miary obrazujące jakość otrzymanych modeli stworzonych za pomocą regresji liniowej wielu zmiennych oraz sztucznych sieci neuronowych

Model	Sieci neuronowe		Regresja liniowa	
	MSE	R^2	MSE	R^2
Model 1	1325,3974151	0,8953	11383,970215	0,53785
Model 2	199,96591289	0,93843	228,82667242	0,84821
Model 3	47,594291589	0,97586	54,134559579	0,96954
Model 4	1 325 349,3979	0,18734	522 782,66346	0,49732
Model 5	592,33779908	0,9593	5014,5613873	0,93893

Źródło: opracowanie własne.

Na uzyskanych danych zastosowano techniki eksploracji danych w celu ustalenia możliwości stworzenia spójnych modeli opisujących zebraną informację oraz wygenerowania modeli dobrze opisujących badane modele wyceny produktów ubezpieczeniowych. Wykorzystane zostały następujące techniki:

- liniowe i nieliniowe modele regresji wielu zmiennych,
- algorytmy programowania genetycznego,
- sztuczne sieci neuronowe.

Zastosowanie trzech różnych technik konstrukcji modeli było celowym zabiegiem badawczym mającym określić, jak poszczególne techniki będą różnić się skutecznością dla zebranych danych.

W rezultacie zastosowania powyżej wymienionych metod analitycznych na danych zagregowanych w kroku poprzednim stworzono szereg przybliżonych modeli wyceny produktów ubezpieczeniowych. Tabele 5 i 6 pozwalają na porównanie średnich błędów kwadratowych (MSE) i współczynnika korelacji (R^2) dla poszczególnych modeli oraz metod ich otrzymywania[63].

63. Miary policzono dla zbioru testowego składającego się z sumy danych treningowych oraz walidacyjnych.

Tabela 6. Miary obrazujące jakość otrzymanych modeli stworzonych za pomocą programowania genetycznego

Model	MSE	R^2	Liczba programów
Model 1	1154,88342	0,95277	16 060 958
Model 2	187,4931	0,9789	11 364 582
Model 3	55,10202	0,9615	11 390 375
Model 4	252 719,578125	0,8690	11 195 494
Model 5	8915,48242	0,9053	10 294 921

Źródło: opracowanie własne.

Z wyjątkiem modelu 5 lepsze rezultaty otrzymane zostały przy zastosowaniu metody programowania genetycznego. Dobre wyniki daje także zastosowanie sztucznych sieci neuronowych. Mimo to przewagą programowania genetycznego jest otrzymanie analitycznej postaci modelu (przykład dla modelu 5 zamieszczono na Rysunku 6).

W przypadku przebadanej próby modeli zastosowana metoda okazała się wysoce skuteczna. Ogólne wnioski można by jednak wyciągać dopiero po przebadaniu większej liczby źródeł.

Rysunek 6. Fragment kodu modelu otrzymanego za pomocą programowania genetycznego

```
f[0]-=optionalriskcbzogoptprzepiecie;  f[0]*=optionalriskcbzogoptszyby;
f[0]-=-1.5865401029586795f;  f[0]=Math.sin(f[0]);
tmp=f[1]; f[1]=f[0]; f[0]=tmp;  f[0]-=ogsumumbezpieczenia;
f[0]/=f[0];  f[1]-=f[0];
f[0]+=f[0];  f[0]+=f[0];
f[0]*=f[0];  f[0]/=f[1];
tmp=f[1]; f[1]=f[0]; f[0]=tmp;
f[0]-=f[1];  f[0]+=ogsumaruchomosci;
f[0]+=ogsumumbezpieczenia;  f[0]+=f[1];
f[0]=Math.sqrt(f[0]);  f[0]/=-0.25369691848754883f;
f[0]/=0.6357779502868652f;  f[0]*=optionalriskstzogsumaszyby;
f[0]+=ogsumubudowle;  f[0]+=ogsumaruchomosci;
f[0]+=ogsumumbezpieczenia;  f[0]+=ogsumaruchomosci;
f[0]+=ogsumumbezpieczenia;  f[0]*=optionalriskcbzogoptpowodz;
f[0]+=ogsumaruchomosci;  f[0]+=ogsumumbezpieczenia;
f[0]+=f[1];  f[0]=Math.sqrt(f[0]);
```

Źródło: opracowanie własne.

4.2. Zastosowania

Do najistotniejszych zastosowań opracowanej metody oraz otrzymanych za jej pomocą wyników można zaliczyć:

- monitoring rynku,
- portale ze zbiorczymi ofertami,
- alternatywny model interoperacyjności,
- cele badawczo-naukowe.

Najciekawszym zastosowaniem wydaje się możliwość automatycznego monitorowania rynku. Poprzez porównywanie modeli w czasie otrzymać można zarówno obraz zmian w skali całego rynku, jak również w skali konkretnych firm. 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, ile dalsze przetworzenie i refleksja nad otrzymaną wiedzą.

Podsumowanie

W niniejszym artykule przedstawiamy rezultaty prac nad prototypem metody i narzędziami ją wspierającymi do realizacji zadania ekstrakcji wiedzy w postaci modelu szacowania wysokości składki ubezpieczeniowej. Omawiamy rezultaty przeprowadzonej analizy źródeł poszukiwanej wiedzy, jak również prezentujemy rozwiązania innych, zbliżonych pod względem technologicznym zadań. Osobnym problemem stanowiącym element zaproponowanego rozwiązania jest dostęp do informacji pochodzącej ze źródła internetowego oraz zagwarantowanie możliwości wszechstronnego przetworzenia pochodzącej z niego informacji. Przeprowadzone badania wskazują, że przy akceptacji określonego poziomu błędu za pomocą wybranych metod i narzędzi analitycznych możliwe jest zbudowanie modeli wyceny. Wielkość błędu zależy od zastosowanych metod oraz charakteru danych w źródle. Wyniki uzyskane przez prototypowe rozwiązanie są zachęcające, choć jeszcze niezadowalające, dlatego konieczne są dalsze prace nad udoskonaleniem części analitycznej.

Wykaz źródeł:

Abramowicz W., Stolarski P., Węcel K., *Ontologie jako narzędzie budowy modeli w ubezpieczeniowych systemach informacyjnych – modelowanie ryzyka oraz produktów cz. 2*, „Wiadomości Ubezpieczeniowe", nr 01/2011, s. 117–137.

Adelberg B., *Nodose – a tool for semi-automatically extracting structured and semistructured data from text documents*, ACM SIGMOD International Conference on Management of Data, s. 283–294, 1998.

Alvarez M., Raposo J., Pan A., Cacheda F., Bellas F., Carneiro V., *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.

Anderson D., et al., *The Practitioner's Guide to Generalized Linear Models*, „Casualty Actuarial Society Study Note", May 2005.

Arabas J., *Wykłady z algorytmów ewolucyjnych*, WNT, Warszawa 2001.

Arasu A., Garcia-Molina H., *Extracting structured data from web pages*, 2003 ACM SIGMOD International Conference on Management of Data, s. 337–348, 2003.

Arocena G.O., Mendelzon A.O., *Weboql: Restructuring documents, databases, and webs*, 14th International Conference on Data Engineering, 1998.

Azavant F., Sahuguet A., *Bulding light-weight wrappers for legacy web data sources using w4f*, 25th International Conference on Very Large Data Bases, 1999.

Berry M.J.A., Linoff G.S., *Mastering Data Mining. The Art and Science of Customer Relationship Management*, Wiley Computer Publishing, New York 2000, s. 494.

Björn A., *Comparison of Some Methods to Fit a Multiplicative Tariff Structure to Observed Risk Data*, „ASTIN Bulletin International Actuarial Association", 1968, vol. 16, nr 1, Belgia, s. 63–68.

Błaszczyszyn B., Rolski T., *Podstawy matematyki ubezpieczeń na życie*, WNT, Warszawa 2004.

Bramer M., *Principles of Data Mining*, Springer, London 2007.

Brown B.Z., Schmitz M.C., *Study Note on Deductibles*, „CAS Study Note", July 2006.

Chawathe S., PapakonstantinouY., Ullman J.D., Garcia-Molina H., Ireland K., Hammer J., Widom J., *The TSIMMIS project: Integration of heterogeneous information sources*, 10th Meeting of the Information Processing Society of Japan, s. 7–18, 1994.

Dionne G., Vanasse C., *A generalization of actuarial automobile insurance rating models. The negative binomial distribution with a regression component*, „Actuarial Bulletin", nr 19/1989, s. 199–212.

Eikvil L., *Information extraction from world wide web – a survey*, raport instytutowy, 1999.

Fraser A.S., *Simulation of genetic systems by automatic digital computers*, „J. Biol. Sci.", nr 10/1957, s. 484–499

Gadomski A.M., *ENEA's e-paper*, http://erg4146.casaccia.enea.it/wwwerg26701/gad-dict.htm, (20.11.2011).

Gerber H.U., *Life Insurance Mathematics*, Springer-Verlag, 1997 (3 wyd.).

He H., Meng W., Yu C., Wu Z, *Wise-integrator: A system for extracting and integrating complex web search interfaces of the deep web*, Proceedings of 31st International Conference on Very Large Data Bases, 2005.

Kaczała M., *Internet jako instrument dystrybucji ubezpieczeniowej*, praca doktorska, UEP 2006.

Kaczmarek T., *Integracja danych z głębokiego Internetu dla potrzeb analizy otoczenia przedsiębiorstwa*, praca doktorska, Akademia Ekonomiczna w Poznaniu, 2006.

Knoblock C.A., Minton S., Ambite J.L., Ashish N., Muslea I., Philpot A.G., Tejada S.,*The ariadne approach to web-based information integration*, „International Journal of Cooperative Information Systems", nr 10(1)/2001, s. 145–169.

Koza J.R., *Genetic Programming II: Automatic Discovery of Resuable Programs*, The MIT Press, 1994.

Kushmerick N., *Wrapper induction for information extraction*, praca doktorska, University of Washington, 1997.

Mahler H.C., Dean C.G., *„Credibility", Chapter 8 in Foundations of Casualty Actuarial Science 4th ed.*, Arlington, VA: Casualty Actuarial Society, 2001.

McCallum A., Cohen W.W., *Information extraction from the world wide web*, tutorial, 2002.

McCulloch W.S., Pitts W.H., *A logical calculus of the ideas immanent in nervous activity*, „Bulletain of Mathematical Biophysics", nr 5/1943, s. 115–133.

Miller G.F., Todd P.M., Hagde S.U., *Designing neural networks using genetic algorithms*, Proc. of the 3rd Int. Conf. on Genetic Algorithms and Their Applications, [red.] J.D. Schaffer, M. Kaufmann, s. 379–384, San Mateo, USA 1989.

Montatna D., Davis L., *Training feedforward neural networks using genetic algorithms*, Proc. of the 11th Int. Conf. on Artificial Intelligence, Morgan Kaufmann, s. 762–767, USA 1989.

Mukkamala S., Sung A.H., Ribeiro B., Vieira A.S., Neves J.C., *Model Selection and Feature Ranking for Financial Distress Classification*, Proc. of 8th Int. Conf. on Enterprise Information Systems (ICEIS), 2006.

Quinlan J.R., *Induction of Decision Trees*, „Machine Learning", nr 1/1986, Morgan Kaufmann, s. 81–106.

Riley J.G., *Competition with Hidden Knowledge*, „Journal of Political Economy"1985, vol. 93, nr 5, s. 958–976.

Salam R., *Estimating the Cost of Commercial Airlines-Catastrophes A Stochastic Simulation Approach*, CAS Forum, Winter 2003.

Shestakov D., Bhowmick S.S., Lim E., *Deque: querying the deep web*, „Data Knowl. Eng.", nr 52(3)/2005, s. 273–311.

Werner G., Modlin C., *„Basic Ratemaking" 4th ed.*, Casualty Actuarial Society, 2010.

Ontologies as a tool for constructing models in insurance information systems – extraction of insurance knowledge from internet sources

Extraction of insurance pricing models from internet sources consists in constructing a representation of the source and characteristic features of the extracted model used to determine relationships between values of variables which are independent of the model, and result from the above-mentioned characteristic features, and the insurance premium amount. Our aim is to obtain a model with a possibly small error calculated as a difference between predicted and actual values. At the same time, we optimise the model reconstruction process which consists in minimising the number of questions sent to the source. Using the information extraction, we actually do not just make use of the database made available by the website, but rather try to identify, on the basis of the data collected, the algorithm (parameters and function form) according to which the premium amount is calculated. Therefore, in contrast to extraction of information from deep internet sources, in the approach proposed we deal with extraction of knowledge. In the part presented, we discuss in detail the method for reconstructing the model, the challenges it involves, and demonstrate a tool which supports its application. Furthermore, we present the results obtained and discuss the scope of application.

PROF. DR HAB. WITOLD ABRAMOWICZ kieruje Katedrą Informatyki Ekonomicznej na Uniwersytecie Ekonomicznym w Poznaniu, prowadzi cykl konferencji Business Information Systems odbywających się co roku.

PIOTR STOLARSKI jest pracownikiem i doktorantem Uniwersytetu Ekonomicznego w Poznaniu, w Katedrze Informatyki Ekonomicznej. Absolwent Wydziału Ekonomii Akademii Ekonomicznej w Poznaniu oraz Wydziału Prawa i Administracji UAM.

DR KRZYSZTOF WĘCEL jest adiunktem w Katedrze Informatyki Ekonomicznej Uniwersytetu Ekonomicznego w Poznaniu.