WOJSKOWA AKADEMIA TECHNICZNA

BIULETYN

INSTYTUTU SYSTEMÓW INFORMATYCZNYCH

WARSZAWA NR 9/2012

BIULETYN INSTYTUTU SYSTEMÓW INFORMATYCZNYCH

ADRES REDAKCJI
Redakcja Biuletynu Instytutu Systemów Informatycznych
Wydział Cybernetyki Wojskowej Akademii Technicznej
00-908 Warszawa, ul. S. Kaliskiego 2
e-mail: biuletynisi@wat.edu.pl
tel.: (22)683-95-04, fax: (22)683-78-58
http://biuletynisi.wcy.wat.edu.pl

Biuletyn ISI jest czasopismem indeksowanym w bazach:
http://baztech.icm.edu.pl/
http://indexcopernicus.com/

Tłumaczenie i korekta tekstu w j. angielskim: Wojciech Gilewski
Opracowanie stylistyczne w j. polskim: Renata Borkowska

Redakcja techniczna i projekt graficzny okładki: Barbara Fedyna

Wydawca: Instytut Systemów Informatycznych Wydziału Cybernetyki WAT

ISSN 1508-4183

Wersją pierwotną (referencyjną) czasopisma jest wydanie papierowe.

Druk: Remigraf Sp. z o.o., ul. Ratuszowa 11, 03-450 Warszawa

SPIS TREŚCI

Translacja opisów ścieżek klinicznych z postaci GLIF na XPDL zapewniająca interoperacyjność z systemem EHR

G. BLIŹNIUK, T. GZIK, J. KOSZELA
gblizniuk@wat.edu.pl

Instytut Systemów Informatycznych
Wydział Cybernetyki WAT
ul. S. Kaliskiego 2, 00-908 Warszawa

W opracowaniu przedstawiono koncepcję translacji zapisów komputerowo interpretowalnych ścieżek klinicznych z postaci GLIF na XPDL opracowaną w ramach realizacji na Wydziale Cybernetyki WAT w latach 2009–2010 projektu POIG.01.03.01-145/08, dofinansowanego ze środków Programu Operacyjnego Innowacyjna Gospodarka w ramach Europejskiego Funduszu Rozwoju Regionalnego. Ponadto przedstawiono sposób zastosowania otrzymanych w wyniku translacji skryptów XPDL do zapewnienia interoperacyjności informatycznego repozytorium ścieżek klinicznych i systemu elektronicznego rekordu pacjenta (EHR).

Słowa kluczowe: clinical pathways, GLIF, XPDL.

1. Wprowadzenie

Modelowanie ścieżek klinicznych to zagadnienie bardzo złożone i czasochłonne. Trudności wynikają m.in. z bardzo dużej różnorodności decyzji i zdarzeń, jakie mogą mieć miejsce w trakcie trwania leczenia. Ich odwzorowanie w postaci procesu wymaga określenia skończonego zbioru wspomnianych zdarzeń i decyzji oraz dysponowania odpowiednią notacją (językiem) modelowania pozwalającą na przedstawienie tak mało przewidywalnych przebiegów. Na rynku dostępnych jest wiele metod notacji i języków dedykowanych do modelowania procesów. Każda z nich może zostać zastosowana z mniejszym lub większym powodzeniem do modelowania ścieżek klinicznych, m.in. BPMN, XPDL, GLIF.

Wykorzystanie notacji BPMN wiąże się z rozbudową metamodelu BPMN o elementy, które umożliwiają definiowanie oraz wykonywanie procesów w sposób dynamiczny. Istotnym faktem i jednocześnie zaletą jest możliwość bezpośredniego przejścia do zapisu zamodelowanej ścieżki w języku XPDL, co z kolei zapewnia właściwie nieograniczone możliwości przenoszenia definicji między różnymi środowiskami klasy workflow. Metoda GLIF z uwagi na fakt, iż jest dedykowana do definiowania ścieżek klinicznych, może zostać wykorzystana bez wprowadzania żadnych zmian. Zawiera „wbudowany" język wyrażeń (ang. *expression language*) pozwalający na rozszerzanie zamodelowanych przebiegów, a zestaw obiektów, jaki dostarcza, jest wystarczający do opisywania ścieżek klinicznych. Wykorzystanie GLIF zapewnia jednak ograniczone możliwości wykorzystania modeli w różnych środowiskach workflow, co przeczy w pewnym zakresie idei ścieżek klinicznych, która zakłada, że powinna istnieć możliwość swobodnego ich udostępniania i przenoszenia. W związku z powyższym zasadne wydaje się wykorzystanie do modelowania ścieżek klinicznych metody GLIF wraz z translacją do języka XPDL.

2. GLIF

W GLIF ścieżki kliniczne modelowane są na trzech poziomach:

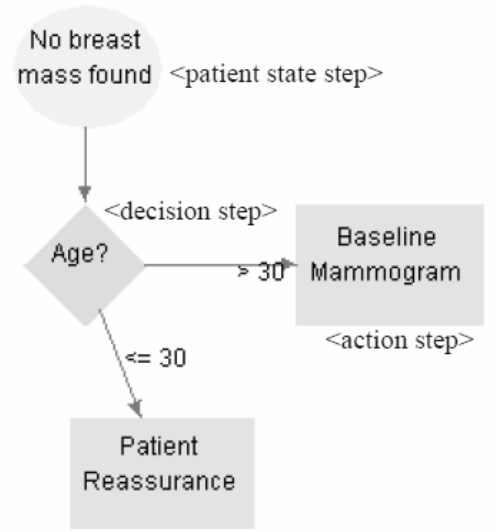
- Conceptual Level (A)
- Computable Level (B)
- Implementable Level (C).

Poziom A obejmuje swoim zakresem przebieg wytycznych klinicznych.
Poziom B stanowi uszczegółowienie Poziomu A, np. kryteria decyzyjne dla poszczególnych punktów decyzyjnych modelu definiowane są na tym poziomie.
Poziom C dotyczy tematów integracji/współpracy z systemami informatycznymi w ramach realizacji modelu.

Metamodel GLIF jest modelem obiektowym składającym się z klas, atrybutów oraz relacji, które są niezbędne do modelowania wytycznych ścieżek klinicznych. Główne klasy metamodelu GLIF:

- klasa Decision_Step – reprezentuje punkt decyzyjny na ścieżce klinicznej
- klasa Action_Step – jest wykorzystywana do zamodelowania akcji wykonywanej w ramach ścieżki. Zawiera zadania, które dzielą się na dwa typy: zorientowane medycznie i implementacyjne
- Branch_Step I Synchronization_Step – umożliwiają modelowanie równoległych przebiegów ścieżek
- Patient_State_Steps – stanowią punkty wejściowe do ścieżki.

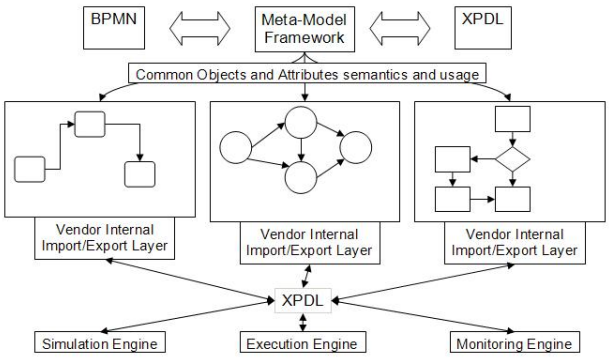
Poniższy rysunek to przykład fragmentu ścieżki medycznej zamodelowanej w GLIF.



Rys. 1. GLIF – przykład, źródło: [1]

3. XPDL

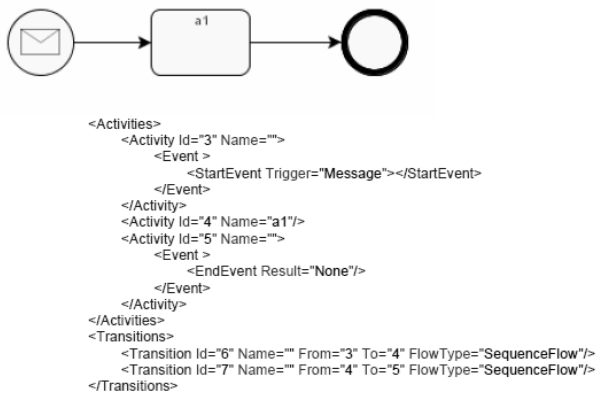
XML Process Definition Language (XPDL) to oparty na XML standard języka opisu procesów. XPDL został stworzony z myślą o wymianie definicji procesów pomiędzy różnymi aplikacjami. Stanowi mechanizm wymiany, który polega na wyeksportowaniu definicji procesu do XPDL w jednej aplikacji, przesłaniu jej i zaimportowaniu w innej aplikacji (rysunek 2).



Rys. 2. Koncepcja wymiany definicji procesów, źródło: [2]

Proces biznesowy w XPDL rozumiany jest jako przepływ prac/zadań (workflow) i opisywany jest w oparciu o metamodel zawierający potrzebne do zdefiniowania procesu obiekty. Metamodel określa logikę i sposób opisywania procesów, który jest zgodny z semantyką notacji BPMN.

Poniżej przykład definicji procesu w języku XPDL:



Rys. 3. XPDL – przykład, źródło: [2]

4. Translacja GLIF – XPDL

Koncepcja translacji GLIF – XPDL (rysunek 4) zakłada, iż definicje wytycznych ścieżek klinicznych tworzone będą zgodnie ze standardem GLIF w dedykowanym do tego celu narzędziu. Definicje generowane będą również do postaci plików XML, które będą stanowiły podstawę do transformacji definicji ścieżki do postaci zgodnej z XPDL. Język XPDL zapewnia możliwość systemowej realizacji ścieżek z wykorzystaniem silnika workflow. Ścieżki zapisane w języku XPDL będą mogły być również przenoszone pomiędzy różnymi (wykorzystującymi ten standard) środowiskami wykonawczymi.

Rys. 4. Translacja GLIF – XPDL

Warto zaznaczyć fakt, iż uzupełnienie przedmiotowego mechanizmu o możliwość generowania definicji ścieżek z postaci XPDL do postaci GLIF zapewniłoby znaczącą „swobodę" w sposobie modelowania wytycznych ścieżek klinicznych i wyborze środowisk definicyjnych oraz wykonawczych. Wiąże się to jednak z koniecznością zdefiniowania sposobu mapowania elementów języka XPDL na elementy języka GLIF, których on nie uwzględnia, np. zdarzenia, elementy pool i lane, artefakty – stanowi to poważne ograniczenie, które znacząco utrudnia i ogranicza budowę rozwiązania informatycznego wspierającego elastyczne definiowanie wytycznych ścieżek klinicznych.

5. Podstawowe elementy (klasy) GLIF

- **Guideline_Collection** – klasa zawierająca listę ścieżek klinicznych
- **Algorithm** – element zawierający listę kroków procesu (steps)

```
<cpr:Algorithm cpr:name="Test"
rdf:about="http://www.skg.pl/cpr/00003">
<cpr:first_step
rdf:resource="http://www.skg.pl/cpr/00004"/>
<cpr:steps
rdf:resource="http://www.skg.pl/cpr/00004"/>
...
</cpr:Algorithm>
```

- **Patient_State_Step** – element stanowiący punkt wejściowy do ścieżki
  - **next_step** – kolejny krok procesu
  - **patient_state_description** – opis stanu ścieżki

```
<cpr:Patient_State_Step
cpr:new_encounter="False"
cpr:name="BADANIE LEKARSKIE"
rdf:about="http://www.skg.pl/cpr/00004">
<cpr:patient_state_description
rdf:resource="http://www.skg.pl/cpr/00005"
/>
<cpr:next_step
rdf:resource="http://www.skg.pl/cpr/00006"
/>
</cpr:Patient_State_Step>
```

- **Decision_Step** – reprezentuje punkt decyzyjny na ścieżce klinicznej
  - **options –** reprezentuje możliwe opcje decyzyjne

```
<cpr:Decision_Step
cpr:automatic_decision="False"
cpr:name="Po TSH"
rdf:about="http://www.skg.pl/cpr/00007">
<cpr:options
rdf:resource="http://www.skg.pl/cpr/00008"
/>
<cpr:options
rdf:resource="http://www.skg.pl/cpr/00011"
/>
<cpr:options
rdf:resource="http://www.skg.pl/cpr/00014"
/>
<cpr:options
rdf:resource="http://www.skg.pl/cpr/00017"
/>
</cpr:Decision_Step>
```

- **Action_Step** – element wykorzystywany do modelowania akcji możliwych do realizacji w ramach ścieżki, zawiera zadania, które dzielą się na dwa typy: zorientowane medycznie i implementacyjne

```
<cpr:Action_Step cpr:name="TSH"
rdf:about="http://www.skg.pl/cpr/00006">
<cpr:next_step
rdf:resource="http://www.skg.pl/cpr/00007"
/>
</cpr:Action_Step>
```

- **Branch_Step, Synchronization_Step** – elementy umożliwiające modelowanie równoległych przebiegów ścieżek.

```
<cpr:Branch_Step cpr:name="Branch"
rdf:about="http://www.skg.pl/cpr/00063">
<cpr:branches
rdf:resource="http://www.skg.pl/cpr/00066"/>
<cpr:branches
rdf:resource="http://www.skg.pl/cpr/00078"/>
</cpr:Branch_Step>
```

6. Podstawowe elementy XPDL

- **Package** – element grupujący elementy procesów

```
<Package
xmlns:xpdl2=http://www.wfmc.org/2008/XPDL2
.1 xmlns:cpr="http://www.skg.pl/cpr">
```

```
<PackageHeader>
<XPDLVersion>2.1a</XPDLVersion>
<Vendor>Sybase PowerDesigner</Vendor>
<Created>15 kwietnia 2010
13:04:30</Created>
</PackageHeader>
...
</Package>
```

- **Application** – reprezentuje narzędzia i aplikacje używane w ramach realizacji procesu
- **WorkflowProcess** – element reprezentujący proces/podproces

```
<WorkflowProcesses>
<WorkflowProcess
Id="http://www.skg.pl/cpr/00003"
Name="Test">
<Activities>
...
</Activity>
...
</WorkflowProcess>
</WorkflowProcesses>
```

- **Activity** – jest podstawowym elementem procesu. Czynności w ramach procesu połączone są za pomocą przejść. XPDL określa trzy typy czynności: Route, Implementation, BlockActivity

```
<Activities>
<Activity Id="-1" Name="Start">
<Event>
<StartEvent />
</Event>
</Activity>
<Activity Id="-2" Name="End">
<Event>
<EndEvent />
</Event>
</Activity>
</Activity>
```

- **Transition** – element łączący elementy procesu.

```
<Transitions>
<Transition Id="2"
From="http://www.skg.pl/cpr/00022" To="-
2" />
<Transition Id="3"
From="http://www.skg.pl/cpr/00024" To="-
2" />
...
</Transitions>
```

- **Participant** – definiuje role w procesie
- **Pool**, **Lane** – elementy umożliwiające definiowanie odpowiedzialności w procesie

```
<Pools>
<Pool Id="http://www.skg.pl/cpr/00002"
Name="Tarczyca"
Process="http://www.skg.pl/cpr/00003">
<Lanes>
<Lane Id="http://www.skg.pl/cpr/00002"
Name="Tarczyca" />
</Lanes>
</Pool>
</Pools>
```

- **MessageFlow** – element zapewniający odwzorowanie mechanizmów komunikacji pomiędzy odrębnymi obszarami odpowiedzialności
- **Artifact** – pozwala przedstawić „dodatkowe" informacje związane z procesem – artefakty.
- **Groups** – element grupowanie elementów procesu
- **DataObjects** – reprezentacja elementów przetwarzanych w ramach procesu
- **Annotations** – notatka uzupełniająca model procesu
- **Gateway** – pozwala wprowadzić dodatkowe elementy decyzyjne i synchronizacyjne
- **Event** – przedstawia różnego rodzaju zdarzenia, które mogą wystąpić w trakcie realizacji procesu.

```
<Activity Id="-1" Name="Start">
<Activity Name="zastosowanie metforminy
w dalszym leczeniu" Id="36" />
</Activity>
```

7. Mapowanie GLIF – XPDL

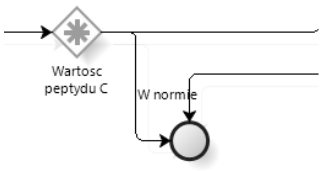
W implementacji transformaty można założyć następujący sposób mapowania elementów GLIF na elementy XPDL:

- Action_Step –> *Activities*
- Patient_State_Step – > DataObject + Activity połączone Association
- Decision_Step –> Gateway XOR
- Branch_Step –> Gateway AND
- Guideline –> *Pool*
- Decision_Option –> *Transition*
- Algorithm –>*WorkflowProcess*

8. Interoperacyjność z systemem EHR

Przedstawione powyżej mechanizmy translacji pomiędzy GLIF i XPDL również zostały wykorzystane do zapewnienia interoperacyjności [1], [4] pomiędzy komputerowym repozytorium ścieżek klinicznych i systemem EHR w wersji OpenMRS[1]. Na rysunkach 1 i 2 z opracowania [5] przedstawiono sposób widzenia systemu EHR, jako systemu zewnętrznego w stosunku do systemu repozytorium ścieżek klinicznych i współpracującego z tym repozytorium za pomocą interfejsu zapewniającego ich interoperacyjność. W projektowych pracach implementacyjnych określono szczegółowo zakresy danych przekazywanych pomiędzy repozytorium ścieżek klinicznych i EHR, co znalazło swój wynik w szczegółowo określonych zakresach danych dla poszczególnych składowych zbiorów D_{E1} i D_{E2}. Dzięki temu umożliwiono skuteczne uruchomienie mechanizmów interoperacyjności pomiędzy systemem ścieżek klinicznych i przykładowym systemem EHR, z zachowaniem warunków przedstawionych w niniejszym opracowaniu.

Na rysunku poniżej przedstawiony jest przykład modelu bramki logicznej dla węzła decyzyjnego w ścieżce klinicznej.



Rys. 5. Przykład bramki logicznej, źródło: [6], [7]

Dla przedstawionej bramki logicznej zaproponowano skrypt XPDL z odpowiednimi zapisami, dzięki czemu maszyna workflow realizująca instancję procesu na podstawie tego skryptu potrafi wspierać działanie bramki logicznej, zgodnie z jej strukturą przedstawioną na rysunku 5:

```
<Activity Id="d3e80159-34e6-44b5-aafe-
d6a1787253fc" Name="Wartosc peptydu C">
        <Description />
        <Route GatewayType="Complex" />
        <Documentation />
        <ExtendedAttributes/>
      <NodeGraphicsInfos>
          <NodeGraphicsInfo
ToolId="BizAgi_Process_Modeler"
Height="40" Width="40" BorderColor="-
5855715" FillColor="-52">
            <Coordinates XCoordinate="282"
YCoordinate="86" />
          </NodeGraphicsInfo>
        </NodeGraphicsInfos>

<IsForCompensationSpecified>false</IsForCom
pensationSpecified>
      </Activity>
```

Fragment źródłowego skryptu XPDL, źródło: [6], [7]

W celu umożliwienia współpracy ścieżki klinicznej z systemem EHR powyższy skrypt został rozszerzony o element *ExtendedAttributes* z odpowiednią wartością.

```
<Activity Id="d3e80159-34e6-44b5-aafe-
d6a1787253fc" Name="Wartosc peptydu C">
        <Description />
        <Route GatewayType="Complex" />
        <Documentation />
        <ExtendedAttributes>
              <ExtendedAttribute Name="EHR"
Value="1.0"/>
        </ExtendedAttributes>
        <NodeGraphicsInfos>
          <NodeGraphicsInfo
ToolId="BizAgi_Process_Modeler"
Height="40" Width="40" BorderColor="-
5855715" FillColor="-52">
            <Coordinates XCoordinate="282"
YCoordinate="86" />
          </NodeGraphicsInfo>
        </NodeGraphicsInfos>

<IsForCompensationSpecified>false</IsForCom
pensationSpecified>
      </Activity>
```

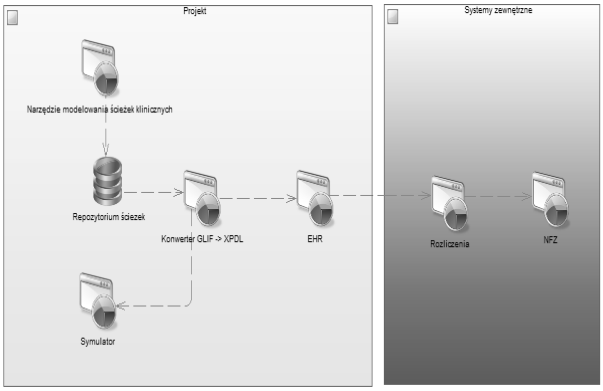
Fragment skryptu XPDL rozszerzonego w elemencie ExtendedAttribute, źródło: [6], [7][2]

Dzięki temu rozszerzeniu w bramce logicznej, zapisanej w skrypcie XPDL, uzyskuje się możliwość współpracy z zewnętrznym systemem EHR. W przyjętej koncepcji skryptów XPDL wartości badań umieszczane są w elemencie *ExtendedAttributes*. Atrybut *Name* ustawiany jest na wartość EHR oznaczającą źródło pochodzenia wyniku badania, natomiast atrybut *Value* przechowuje rzeczywisty wynik badania. Warto nadmienić, że ze względu na

[1] http://openmrs.org/wiki/OpenMRS

[2] W tym przypadku przyjęto przedział wartości peptydu C według skali 0,7-2,0 mcg/l.

przyjętą logikę mapowania elementów ścieżki klinicznej w systemie EHR, w odpowiednich skryptach XPDL nie należy zmieniać nazw bramek logicznych, czyli elementu *Activity* i atrybutu *Name*. Dodatkowym uwarunkowaniem przyjętych rozwiązań jest brak pełnej implementacji standardu grupowania wyników badań szpitalnych i laboratoryjnych LOINC (patrz: [8]), co wynikało z ograniczeń zakresu projektu POIG.01.03.01-00-145/08. Kolejnym zagadnieniem istotnym dla przedstawionych w tym miejscu rozważań było uzupełnienie zapisów w skrypcie XPDL o wartości umożliwiające identyfikację według kodów międzynarodowej klasyfikacji procedur medycznych ICD-9 oraz międzynarodowych kodów klasyfikacji chorób i innych problemów zdrowotnych ICD-10 [8]. Było to ważne ze względu na konieczność standaryzacji zapisu tych danych w systemie EHR, a także dla skutecznego rozliczania świadczeń medycznych, zapisanych w systemie EHR na podstawie opisów ścieżek klinicznych, przechowywanych w ich elektronicznym repozytorium, o czym mowa w dalszej części opracowania.



Rys. 6. Schemat przepływów danych dla ich konwersji, źródło: [7]

Na schemacie przedstawionym na rysunku 6 zobrazowano dodatkowe przepływy danych z systemu EHR do systemu rozliczeń, realizowanych zgodnie z wytycznymi NFZ, dotyczących klasyfikacji wykonanych usług medycznych zgodnie z regułami tzw. jednorodnych grup pacjentów (JGP). W celu umożliwienia realizacji mechanizmów tych rozliczeń, w pracy [6] zaproponowano odpowiednią modyfikację skryptów XPDL i odpowiednie odniesienia do standardów słowników ICD-9 i ICD-10. Przykład dla aktywności z listą ICD-9 został przedstawiony poniżej.

```
<Activity Id="newpkg1_wp1_act3"
Name="Podanie leku trombolitycznego trzeciej
generacji">
        <Implementation>
        <No/>
                </Implementation>
                <ExtendedAttributes>
                        <ExtendedAttribute
Name="ICD9" Value="99.103"/>
                </ExtendedAttributes>
</Activity>
```

Fragment skryptu XPDL uwzględniającego kod ICD-9, źródło: [6]

W aktywności przedstawionej na powyższym listingu jej nazwa jednoznacznie odzwierciedla wykonaną procedurę medyczną. Należy pamiętać, że po jej zmianie przypisanie jej do odpowiedniego kodu ICD-9 mogłoby okazać się niemożliwe. Z tego powodu w docelowych implementacjach należy zapewnić odpowiednie powiązania nazw i kodów poszczególnych pozycji słownika ICD-9.

Poniżej przedstawiono przykład dla aktywności z wykorzystaniem przykładowej wartości ze słownika ICD-10.

```
<Activity Id="newpkg1_wp1_act14"
Name="Szczegółowe badania">
        <Implementation>
                <No/>
        </Implementation>
        <ExtendedAttributes>
                <ExtendedAttribute
Name="ICD10 " Value=" E10.5"/>
        </ExtendedAttributes>
</Activity>
```

Fragment skryptu XPDL uwzględniającego kod ICD-10, źródło: [6]

W tym przypadku również należy pamiętać o konieczności odpowiedniego wiązania nazw i kodów poszczególnych pozycji słownika ICD-10. Pozycje skryptu XPDL, dla których nie przypisano żadnego kodu, nie mają znaczenia dla zapisu historii leczenia pacjenta i rozliczeń usług medycznych. Pełnią one wtedy funkcję wyłącznie informacyjną.

9. Podsumowanie

W zagadnieniu przedstawionym w niniejszym opracowaniu kluczowe było opracowanie odpowiednich mechanizmów translacji pomiędzy różnymi standardami i formatami opisu procesów workflow w zastosowaniach medycznych. Dodatkową trudnością była

konieczność uzyskania interoperacyjności pomiędzy systemem repozytorium ścieżek klinicznych i przykładowym systemem EHR.

Wobec powyższego przyjęto dość logiczną koncepcję zastosowania języków znacznikowych pochodzących z rodziny XML lub koncepcyjnie do nich podobnych po to, aby możliwe było efektywne zdefiniowanie i zaimplementowanie reguł interfejsowych dla poszczególnych komponentów systemu ścieżek klinicznych.

Przedstawione powyżej przykłady rozwiązań ilustrują istotę przyjętych rozwiązań i mogą stanowić sugestię dla rozwoju przyjętej koncepcji mechanizmów konwersji oraz metod osiągania interoperacyjności systemów.

10. Bibliografia

[1] Guideline Interchange Format Technical Specification, 2004.

[2] Workflow Management Coalition, *XML Process Definition Language Specification*, październik, 2008.

[3] G. Bliźniuk, „O kilku warunkach interoperacyjności systemów informacyjnych i informatycznych”*, Biuletyn Instytutu Systemów Informatycznych*, Nr 3, str. 13–18 (2009).

[4] G. Bliźniuk, ”Thing about Some Assuring Interoperability of Information and Information Technology Systems Conditions”, *Polish Journal of Environmental Studies*, Vol. 18, No. 3B, 30–34 (2009).

[5] G. Bliźniuk, „Określenie przydatności standardów BPMN, GELLO, UML, OCL, XML, HL7 i ich wybór dla modelu repozytorium. Kontekst zapewnienia interoperacyjności”, w: *Raport końcowy projektu* POIG.01.03.01-00-145/08, str. 311–318, WAT, Warszawa, 2010.

[6] I. Iwicki, *Implementacja mechanizmów zapewniających rozliczanie usług medycznych z wykorzystaniem opisu ścieżek klinicznych*, praca magisterska pod kierunkiem G. Bliźniuka, Wydział Cybernetyki WAT, Warszawa, 2010.

[7] P. Giętkowski, *Implementacja mechanizmów zapewniających interoperacyjność systemów EHR i systemów ścieżek klinicznych*, praca magisterska, Wydział Cybernetyki WAT, Warszawa, 2010.

[8] G. Bliźniuk, „Ranking inicjatyw standaryzacyjnych oraz standardów kluczowych dla opisu wytycznych i ścieżek klinicznych”, w: *Metody i narzędzia projektowania komputerowych systemów medycznych*, str. 52–60, Vizja Press & IT, Warszawa, 2009.

[9] R. Bronowski, „Inicjatywy standaryzacyjne z dziedziny systemów wspomagających podejmowanie decyzji klinicznych”, w: *Raport końcowy projektu* POIG.01.03.01-00-145/08, str. 101–110, WAT, Warszawa, 2010.

[10] J. Dytfeld, „Medyczny opis ścieżki klinicznej dla cukrzycy”, w: *Raport końcowy projektu* POIG.01.03.01-00--145/08, str. 601–604, WAT, Warszawa, 2010.

[11] T. Gzik, „Wykonanie metamodelu wytycznej ścieżki klinicznej”, w: *Raport końcowy projektu* POIG.01.03.01-00--145/08, str. 503–506, WAT, Warszawa, 2010.

[12] J. Koszela, „Opracowanie oceny przydatności metod standaryzacji opisu planu wykonywania instancji procesów działalności w kontekście wytycznych i ścieżek klinicznych”, w: *Raport końcowy projektu* POIG.01.03.01-00-145/08, str. 187–191, WAT, Warszawa, 2010.

[13] M. Lignowska, „Uruchomienie narzędzia badań symulacyjnych w ramach narzędzia badań efektywnościowych”, w: *Raport końcowy projektu* POIG.01.03.01-00--145/08, str. 659–672, WAT, Warszawa, 2010.

[14] T. Nowicki, „Miary efektywności informacyjnej w opisach wytycznych i ścieżek klinicznych”, w: *Raport końcowy projektu* POIG.01.03.01-00-145/08, str. 221–231, WAT, Warszawa, 2010.

[15] S. Palicki, M. Dobkowski, P. Dąbrowski, „Specyfikacja techniczna transformaty definicji ścieżek klinicznych”, dokumentacja wykonana w ramach zadania 4. projektu POIG.01.03.01-00-145/08

[16] D. Tukaj, „Ścieżki kliniczne – co to jest i jak je wytyczać?”, *Ogólnopolski Przegląd Medyczny*, Nr 9–10, str. 72–74 (2005).

[17] T. Zdrojewski, „Opracowanie oceny adekwatności doboru elektronicznych źródeł wiedzy medycznej pod kątem reprezentatywności opisów wytycznych i ścieżek klinicznych”, w: *Raport końcowy projektu* POIG.01.03.01-00-145/08, str. 243–250, WAT, Warszawa, 2010.

[18] Strona organizacji Workflow Management Coalition, http://www.wfmc.org/

Translating descriptions of clinical pathways from GLIF to XPDL that provides interoperability with EHR system

G. BLIŹNIUK, T. GZIK, J. KOSZELA

The article presents the concept of translating the definition of clinical pathways from GLIF to XPDL, which was developed at the Department of Cybernetics WAT in 2009-2010, under POIG.01.03.01-145/08 project, funded by the Operational Programme – Innovative Economy under the European Regional Development Fund. Furthermore, the article shows how to apply resulting XPDL translation scripts to ensure interoperability of information repository of clinical pathways and an electronic health record (EHR).

Keywords: clinical pathways, GLIF, XPDL.

A Model the Process of Overcoming Multizone Protection Stationary Object by an Intruder

G. KONOPACKI
gkonopacki@wat.edu.pl

Institute of Computer and Information Systems
Faculty of Cybernetics, Military University of Technology
Kaliskiego Str. 2, 00-908 Warsaw, Poland

The article examines a model of the process to overcome multi-zone protection for a stationary object (buildings, together with the adjacent area) by a determined passive intruder, which means the intruder is not affecting the active protection system (for both equipment and people) and not intending to stop the action before achieving a protected object. As a tool for describing the actions the intruder was proposed a process of Markov class CD, whose character is presented in the form of analytical equations Chapman – Kolmogorov. The article presents a solution to this system and discusses its practical usefulness.

Keywords: protection of objects, modeling, the process of Markov.

1. Introduction

The need to ensure the safety of various objects, especially the importance of military, political, financial, is becoming more common and already existing results from evolution and the occurrence of new threats due to the brutalization of the methods attackers use and an increase the value of the damage. Thus, it became a standard to equip objects in a more or less complex system of protection, so called. protection systems of objects. Usually, they combine two elements working together: the technical system and physical protection, in which the crucial link is the person.

Nowadays, the rapid development of the technical system is subject to a particular security system, which does not detract from a person of decisive importance in providing effective protection, as he/she makes final decisions based on the reaction of the technical system. Nevertheless, the electronics and computer science play a significant role in the protection systems of objects. The observation area is carried out using CCTV cameras and object access control is controlled by computer systems, because they are actually only capable of uniformly continuous and reliable operation in various climatic conditions at different times of the day and year. With well-developed multimedia technology, it is possible to visualize data and events in the security system and protected objects.

An undeniable advantage of such systems is that they can also be equipped with elements of artificial intelligence, used to analyze data on the state of the protected object, and, most importantly, to detect, locate, and often neutralize the actions of an intruder, that is violating a protection zone of the protected object. The requirements of modern computer security systems are extremely high. Their fulfilment can guarantee only systems having the following properties:

- high technical reliability
- reaction to the materialization of credibility (the implementation) the risks, in particular the reliability of detection and location of the intruder, a threat to the protected object
- a minimum level of occurrence of states of a false alarm and false peace
- ease of signals verification generated by the system
- ease of use
- resistance to sabotage and being destroyed.

Advanced computer security systems offer the possibility of developing proposals for decision in the event of specific threats. It is very important that the realization of risks (particularly for high intensity) may cause different reactions of the person responsible for taking decisions and actions in the event of abnormal signalling facility. Thus, the system offers staff assistance in the form of ”hints” of activities, forcing their specific sequence, documenting decisions and actions, and recalls the actions necessary, but have not yet taken.

In cases where the implementation of a local hazard is mild, this aspect of the system is not so very important. While the realization of the risks over a large area, with their high intensity, requires rapid decision making during the coordination of activities aimed at neutralizing the effects of the implementation and/or managing the rescue operation. In such situation, large amounts of data come to management positions that require their rapid interpretation and treatment through decisions.

An equally important factor as the efficiency and effectiveness of technical security is the reaction of appropriate services generated by the alarm system. Therefore, the security system equipped with a reliable and efficient installation of signalling and notification, and a competent person to respond to the occurrence of hazards provide proper security of the protected object. Construction of computer systems for the protection of buildings must take into account the principles, which show that this system must [Niezabitowska 2010]:

- be closely tailored to the specifics of the protected object and its protected value. This usually determines that equal security systems is not created for two different protected objects, even if characterized by very similar properties, since the same details to facilitate security systems are able to be overcome relatively quickly which, in turn, would result in futility to continue their use
- include an appropriate set of technical devices (sensors, cameras, lights, analyzers), to ensure providers identify when the materialization of risks
- be flexible, that is designed for easy expansion and changes resulting from changes in the protected object and technical development of protective devices, and from changes in intruder techniques and activities used by the tools used for active conservation measures applied against the object
- provide certainty that the predetermined probability of materialization are detected signs of danger (e.g. intruder detection) in such intensity that they can give rise to legitimate concern about the impact on the protected object. Thus, it is clear that if the status of the security object can be expected to reduce, it must also change the security system itself.

There is another expectation directed at the security systems associated with their ”intelligence”. Modern computer security systems must provide protection to develop proposals for decision support for identification of the realization of a particular type of threat. Thus, understood as ”intelligence” artificial intelligence would be used primarily to recognize the state of a ”false alarm” and ”false peace”, the identification of individuals according to their somatic characteristics such as fingerprints, bones, skulls, DNA and other individuating characteristics between individuals and the identification of adverse, or other desired states and events. The use of artificial intelligence, because of its cost, should occur in such cases where the intensity necessary to observe the events is so high that physical protection (people) cannot guarantee a sufficiently high probability of identifying emerging signs of abnormalities or risk realization

Reliability of security systems are strongly associated with the immune system on the prevalence of ”false alarms” and ”false peace”. A false alarm occurs when the protected object and system security measures are not subject to the security system intruder alarm signals. This condition can be caused by defective functioning of technical elements of the system or − at the technically efficient system − occurrence of adverse random events (e.g. storm, accidental activation barrier). False peace − a phenomenon far more dangerous than a false alarm − occurs when there is actual penetration of a protected object or its environment by an intruder, while the security system for various reasons does not respond.

The principal tasks of the vast majority of security systems, objects are oriented to prevent an attacker (sniffer) on the area immediately adjacent to the protected object, and for his intrusion into the area − its location as quickly as possible, secretive surveillance and eventual neutralization.

2. A Model

2.1. The Assumptions

A formal proposal to approach the problem of overcoming the multi-zone intruder protection area of the object will be presented next. This problem will be dealt with using the following assumptions:

1. The area protected facility (facility protection area) creates a number of concentrically located relative to the object

of the protection zones. Protection zones are disjoint and adjacent closed areas, armed with various physical and technical measures, which is another inconvenience for an intruder on the way to a protected object. Examples of protection zones for the object may be [Nowak 2007]:

- *the first zone of protection* − building a fence that can be equipped with sensors that detect attempts by his clambered sufficient safety, or performance of the grid intersection of sap
- *the second zone of protection* − the area of land between the fence and building wall
- *the third zone of protection* − the technical elements of signaling the presence of an intruder at the wall of an object or attempts of overcoming by him external insurances of security of objects
- *the fourth protection zone* − inside the object.

2. Overcoming the security zones by the intruder begins to break the protective barrier of the outer zone of the protected area, then move to the next zone, located directly to the property protected. The way to overcome the protected area depends on the skills and preparation of the intruder and the organization of the protection zones, which means that the following scenarios are possible:
   - intruder leaves the protected area at the break of the outer barrier (first) zone of protection (an accidental intruder or hearing the security system of the protected object)
   - intruder, not paying attention to whether it was detected, possibly overcomes the simplest way (in the shortest time) the protection zone, until they reach the protected object
   - intruder defeats the zone of delay resulting from the search path of transition to the next zone and the possible withdrawal of the zone immediately preceding that in which it is located, to leave the protected area without reaching the protected object
   - intruder moves as before with the fact that he/she can ”jump” (e.g. with the use of means of transport), individually or collectively forward some or all zones, but does not intend to leave the protected area before reaching the protected object (intruder determined).

The intruder waiting in the protection zone can last for a stretch of time, which is the realization of a random variable with distribution depending on the preparation of the intruder, his/her strategy for overcoming the protected area and protection zones to organize, while the transition between zones of protection occurs without loss of time (at $t = 0$).

Due to the practical aspects, this will be processed on a model, which takes into account the conservation area to overcome the object by a determined intruder, i.e., one that is determined to achieve the protected object.

2.2. The Formulation of the Problem

It is assumed that the facility protection area consists of a finite number of $n \in N$ protection zones. Let $S_i, (i = 0,1,2,...,n)$ mean protection zone number i, where S_0 means outer protection zone, while S_n denotes the last zone, near the protected object.

Based on the verbal description of the process of overcoming the facility protection area by the intruder suggests the following formal model based on stochastic process of the CD class (continuous parameter, discrete states) in which the parameter will be the time for a set of states − a collection of protection zones. A graphic illustration of this process is illustrated in Figure 1, which adopted the following designations:

- $\lambda_{i,j}(t)$ – intensity of the transition intruder from the protection zone S_i to zone S_j $(i < j)$
- $\eta_i(t)$ – intensity of the immediate transition of the intruder from the protection zone S_i to zone S_n
- $\mu_{j,i}(t)$ – intensity gradually withdrawing intruder from the protection zone S_j to zone S_i $(i < j$ and $i \neq j)$
- $\xi_i(t)$ – intensity of the immediate withdrawal of the intruder from protection zone S_i to zone S_0 $(i \neq n$ and $i \neq 0)$
- $\mu_{i,i}(t)$ – intensity of an intruder waiting in zone S_i, where:

$$\begin{cases} \mu_{i,i}(t)=1-\left(\eta_i(t)+\xi_i(t)+\sum\limits_{k=i+1}^{n}\lambda_{i,k}(t)+\right. \\ \left.+\sum\limits_{k=0}^{i-1}\mu_{i,k}(t)\right), \qquad i=0,1,...,n\text{-}1, \\ \mu_{n,n}(t)=0. \end{cases} \tag{1}$$

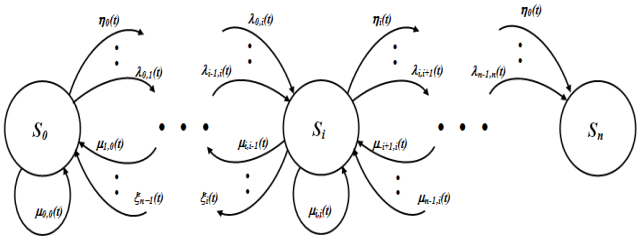


Fig. 1. Graphical presentation of the process of overcoming the multi-zone protection area for the object by an intruder

Examining the process of overcoming the protection area of the object can be treated as a homogeneous Markov process with continuous and discrete parameters in the states. Parameter in this process is time, while the states − as had already been said − the zone of protection. Thus, this process can be presented in the following system of Chapman − Kolmogorov equations:

$$\begin{cases} p_i'(t)=-\left(\sum\limits_{k=0}^{i-1}\mu_{i,kt}(t)+\sum\limits_{k=i+1}^{n}\lambda_{i,k}(t)+\eta_i(t)+\xi_i(t)\right)\cdot \\ \cdot\, p_i(t)+\sum\limits_{k=0}^{i-1}\lambda_{k,i}(t)\cdot p_k(t)+ \\ +\sum\limits_{k=i+1}^{n-1}\mu_{k,i}(t)\cdot p_k(t), \quad i=0,1,...,n-1, \\ p_n'(t)=\sum\limits_{k=0}^{n-1}\left(\lambda_{k,n}(t)+\eta_k(t)\right)\cdot p_k(t), \\ i=n. \end{cases} \tag{2}$$

Determination of probabilities $p_n(t), p_{n-1}(t),..., p_0(t)$, finding the intruder in zone $S_i, (i=0,1,2,...,n)$ requires the solution of this system of equations with the initial conditions defining the probability of finding the intruder in any protection zone. In the general case, the solution of the Chapman − Kolmogorov equations can be very difficult in an analytical way. In such cases, approximate methods are used, using the properties of the process.

In practice, simpler options are used to overcome the facility security area issue by the intruder. This simplified model will be presented below.

2.3. Simplified Model

Model simplified treated further illustrates the often encountered case of the intruder only nudge forward, whose aim is to achieve as soon as possible the protected object. Thus, this process will be only characterized with intensities of transitions between adjacent states, showing the successive zones of protection, the protected object coming closer and the intensities go from any state to the state of the imaging area directly adjacent to the protected object. In the present model of the process, it is assumed that the intensity of the transitions between states are constant (not time-dependent), and that

$$\begin{aligned} &\eta_i>0, \quad i=0,1,...,n-1, \\ &\lambda_{i,i+1}>0, \ i=0,1,...,n-1, \\ &\mu_{i+1,i}=0, \ i=n-2,...,1,0, \\ &\varepsilon_i=0, \quad i=n-1,...,1,0. \end{aligned} \tag{3}$$

Graphical interpretation of the simplified model to overcome the multi-zone protection area by the intruder object is shown in Figure 2.

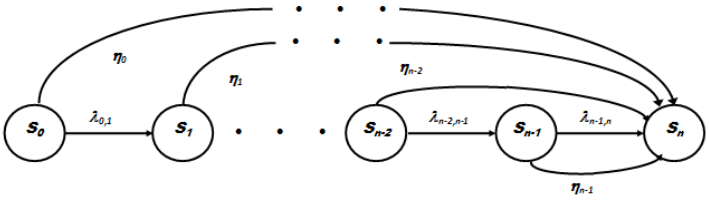


Fig. 2. Graphical presentation of a simplified model of the process of overcoming the multi-zone protection area for the object by an intruder

The stochastic process under consideration can be described by the following system of equations of Chapman − Kolmogorov [Gichman & Skorochod 1968, Norris 1977, Stefanko 2000]:

$$\begin{cases} p_0'(t)=-(\eta_n+\lambda_{0,1})\cdot p_0(t), \\ \qquad i=0, \\ p_i'(t)=-(\eta_i+\lambda_{i,i+1})\cdot p_i(t)+\lambda_{i-1,i}\cdot p_{i-1}(t), \\ \qquad i=1,2,...,n\text{-}1, \\ p_n'(t)=\sum\limits_{k=0}^{n-1}(\lambda_{k,n}+\eta_k)\cdot p_k(t), \\ \qquad i=n. \end{cases} \tag{4}$$

It is assumed that

$$\alpha_i=\eta_i+\lambda_{i,i+1}, \qquad i=0,1,2,...,n-1. \tag{5}$$

With the numerical values set the intensity of transitions between zones of protection can be met with two cases:

- values a_i are different
- not all values a_i are different.

Next, consider the first variant. In the case of the second variant, shown on a method of solving the equations (4) it will be possible to use. At $t_0 = 0$ the intruder can be located in any zone of protection S_i, $(i = 0,1,2,...,n)$ with a probability p_i, $(i = 0,1,2,...,n)$. Therefore, the system of differential equations (4) is solved with the following initial conditions (Cauchy):

$$p_i(t_0) = p_i, \qquad i = 0,1,2,...,n-1,$$
$$\sum_{i=0}^{n} p_i = 1. \tag{6}$$

Applying the Laplace transformation to equations (4) is obtained [Stewart 1994]:

$$p_i(s) = \frac{1}{s+\alpha_i}\left(p_i + \sum_{k=0}^{i-1} p_k \prod_{j=k}^{i-1} \frac{\lambda_{j,j+1}}{s+\alpha_j} \right), \tag{7}$$
$$i = 0,1,2,...,n-1.$$

Applying the appropriate transformations to the system of equations (7) is replaced by the following equations:

$$p_i(s) = \sum_{j=0}^{i} A_{ij} \frac{1}{s+\alpha_j}, i = 0,1,2,...,n-1, \tag{8}$$

where:

$$A_{ij} = \sum_{k=0}^{j} p_k \frac{\prod_{\substack{m=k \\ m\neq i}}^{i} \lambda_{m,m+1}}{\prod_{\substack{l=k \\ l\neq j}}^{i} (\alpha_l - \alpha_j)}. \tag{9}$$

Applying inverse Laplace transformation to equations (8) the following is obtained:

$$p_i(t) = \sum_{j=0}^{i} A_{ij} \cdot e^{-\alpha_j t}, \; i = 0,1,2,...,n-1. \tag{10}$$

Probability $P(t)$ that the intruder will not get in the zone closest to the protected object is equal to:

$$P(t) = \sum_{i=0}^{n-1} p_i(t). \tag{11}$$

Given the expressions (9) and (10) in (11) we finally obtain:

$$P(t) = \sum_{i=0}^{n-1} B_{in} \cdot e^{-\alpha_j t}, \tag{12}$$

where:

$$B_{in} = \left[1 + \sum_{\substack{k=i \\ i\neq n-1}}^{n-2} \prod_{j=1}^{k} \frac{\lambda_{j,j+1}}{\alpha_{j+1} - \alpha_i} \right] \cdot \left[p_i + \sum_{\substack{k=1 \\ i\neq 0}}^{i} p_{i-k} \prod_{j=i-1}^{i-k} \frac{\lambda_{j,j+1}}{\alpha_j - \alpha_i} \right]. \tag{13}$$

From the expressions (12) the density distribution function can be calculated of finding the intruder in any of the protection zones except near the protected object, which is expressed by the following formula:

$$f(t) = \frac{dP(t)}{dt} = \sum_{i-0}^{n-1} \alpha_i \cdot B_{in} \cdot e^{-\alpha_i t}. \tag{14}$$

The mean value of time finding the intruder in any of the protection zones except near the protected object will be equal:

$$ET = \int_0^{\infty} P(t)dt = \sum_{i-0}^{n-1} \frac{B_{in}}{\alpha_i}, \tag{15}$$

and the variance of the time

$$\sigma_t^2 = 2\sum_{i-0}^{n-1} \frac{B_{in}}{\alpha_i^2} - ET^2. \tag{16}$$

3. Conclusion

The practical use of the presented model must be preceded by an estimate of the size occurring in the output, which are the intensity of the protection zones, were created to overcome the storage facility by an intruder, the group. Of course they will depend on the preparation and determination of an intruder on the one hand and the degree of difficulty of overcoming the zone, which in turn depends on its previous facilities and appropriate security measures. Model consideration involves overcoming zones by a determined intruder, but passive, i.e. not affecting destructive to the security measures installed in the zone, and thus not reducing its protective properties. Such situations occur most frequently, and all kinds of intruder impacts on the of security measures in the zone are random.

Despite the high level of generality considered, the model can be used in designing security systems for storage facilities for even approximate estimates of the expected time *ET* to find an intruder outside the immediate neighborhood of the protected storage object, i.e. outside S_n, because it is actually based on the evaluation of the effectiveness of the security system expressed in the following, applied in practice, the relation:

$$ET > T_a + T_m + T_i, \quad (17)$$

where:

- ET – mean time to overcome the protection zones
- T_α – reaction time of the system (alarm call)
- T_m – time of signal arrival to the monitoring centre
- T_i – time necessary for effective intervention.

Bibliography

[1] I.I. Gichman, A.W. Skorochod, *Wstęp do teorii procesów stochastycznych*, PWN, Warszawa, 1968.

[2] W. Feler, *Wstęp do rachunku prawdopodobieństwa*, PWN, Warszawa, 1996.

[3] O. Haggstrom, *Finite markov chains and algorithmic applications*, Chalmers University of Technology, 2001.

[4] G. Konopacki, K. Worwa, „Problem eliminowania fałszywych alarmów w komputerowych systemach ochrony peryferyjnej", *Biuletyn Instytutu Systemów Informatycznych*, Nr 5, 37−46 (2010).

[5] G.F. Lawler, *Introduction to Stochastic Processes*, Chapman & Hall / CRC, 1995.

[6] M. Mitzenmacher, E. Upfal, *Metody probabilistyczne i obliczenia*, WNT, Warszawa, 2009.

[7] E. Niezabitowska (red.), *Budynek inteligentny*, Politechnika Śląska, Gliwice, 2010.

[8] M. Nowak, „Strefy ochrony posesji", www.budujemydom.pl, (2007).

[9] J.R. Norris, *Markov Chains* (Cambridge Series in Statistical and Probabilistic Mathematics), University of Cambridge, United Kingdom, 1997.

[10] A. Papoulis, *Prawdopodobieństwo, zmienne losowe i procesy stochastyczne*, WNT, Warszawa, 1972.

[11] K. Sobczyk, *Stochastyczne równania różniczkowe*, WNT, Warszawa, 1996.

[12] O. Stenflo, "Ergodic theorems for Markov chains represented by iterated function systems", in: *Bulletin Polish Academy of Sciences*, 2000.

[13] W.J. Stewart, *Introduction to the Numerical analysis of Markov Chains*, Princeton, 1994.

Model procesu pokonywania wielostrefowej ochrony obiektu stacjonarnego przez intruza

G. KONOPACKI

W artykule rozpatruje się model procesu pokonywania wielostrefowej ochrony obiektu stacjonarnego (zabudowania wraz z przyległym terenem) przez pasywnego intruza zdeterminowanego, co oznacza intruza nie oddziałującego czynnie na system ochrony (dotyczy zarówno urządzeń, jak i osób) oraz nie zamierzającego przerwać działań przed ociągnięciem celu, tj. obiektu chronionego. Jako narzędzie opisu procesu działań intruza został zaproponowany proces Markowa klasy CD, którego postać analityczną przedstawiono w postaci układu równań Chapmana − Kołmogorowa. W artykule przedstawiono rozwiązanie tego układu i omówiono jego przydatność praktyczną.

Słowa kluczowe: ochrona obiektów, modelowanie, proces Markowa.

Data Visualization While Determining Similarities of Medical Patterns

T. RZEŹNICZAK
tomek.rzezniczak@gmail.com

Institute of Computer and Information Systems
Faculty of Cybernetics, Military University of Technology
Kaliskiego Str. 2, 00-908 Warsaw, Poland

The article presents the concept of using the theory of similarity in the recognition of medical patterns. The aim of the work is to construct a graphical model of disease entity pattern and the state of the patient's health in such a way as to use natural human ability of perception to identify similarities between them. With this approach, the representation of medical patterns can be used to support the diagnosis process of disease entities.

Keywords: data visualization, similarity models, similarity relation, medical diagnostic.

1. Introduction

The job of physicians is based on processing large amounts of medical information describing the patient's health status, on which physicians make decisions and direct the treatment. With the development of science and technology, the number of information sources continues to grow. Physicians now have to their disposal, apart from medical interviews and physical examinations, much more specialized tests. Despite the advanced technological developments, it is still the logical thinking of the physician in conjunction with information collected in many kinds of tests that is the basis of diagnosis. By examining, the physician wants to gather as much information since any of them may affect the diagnosis. At the same time the amount of data increases the difficulty of their analysis, which is the basis for identifying the syndrome.

To set the initial or the final diagnosis, the doctor must critically evaluate the information collected and match them with known disease entities. Given the amount of possible disease entities and the amount of medical data, this task in fact is not an easy one. In addition, many factors that can cause errors, have an impact on its outcome.

Erroneous medical decisions are frequently cognitive – they are errors of reasoning, which are caused by emotions that influence the perception of physicians and their activity [12]. An example could be the expected confirmation of a diagnosis by carefully selecting information or diagnosing just the easier to associate diseases, and forgetting about the rare cases. Besides, no one is able to master the entire medical knowledge, so some errors are due to ignoring the uncertainty associated with the lack of knowledge.

Given the above issues, this work is to initially verify the applicability of visualization methods to support the recognition of disease entities and to conduct treatment. Visualization methods have already been applied in many areas of life such as science, business and media. Examples of their use can be noticed when watching the weather forecast, tracking stock market results or using maps. Visualization is common since the explosion of information forced the search for more effective methods of their processing, and thanks to the innate abilities of human perception, visualization is a very powerful tool. Well-designed visual representation allows to quickly receive information and to analyze more amounts of data by a human than with other methods.

Returning to the medical diagnosis process, the work focuses on the possibility of developing a graphic form of the patient's health condition and patterns of disease entities in order to use the natural abilities of perception to recognize the similarities between them. In addition to the methods of visualization, theory of similarity plays a key role in the work [16], [24], [20], [13]. Similarity models are the basis for the construction of a visualization, they serve as guidelines for the constructing a graphical representation and its evaluation. The tasks rely on finding the optimal visualization, it is one that most effectively supports the diagnosis by comparing the patient's health condition with the disease entity.

2. Similarity Models

As already mentioned, the theoretical basis of the work are based on the analysis of similarity relations. Similarity is the foundation of cognition, it allows the activation of memory according to what we see [14], the categorization of objects [22], decision making or solving new problems based on similar, previously known situations [21]. In the context of psychological similarity between objects we can define it as the mental representation proximity of these objects.

Many models have been created describing the relationship of similarity, among which geometric models dominate. This type of model represents each object as a point in space (usually Euclidean space), and the distance between points corresponds to the similarity of objects. An example of such a model is MDS (Multidimensional Scaling) [11]. Part of the model is a statistical technique, for which the input data are evaluations of similarities or differences between all objects in the model under consideration. The result of the technique is a geometric model representing objects as points in an n-dimensional space.

Formally, MDS can be described as follows: let k be the number of all objects under consideration and n is the number of attributes of each object. Matrix X with a dimension of $k \times n$ will contain the spatial coordinates of the objects, where the row i indicates the coordinates of the object i. However, the difference between objects i and j, will be described by δ_{ij}. The distance in the Euclidean space between objects i and j is defined as the shortest line connecting i with j and takes on the form:

$$d_{ij}(X) = \left(\sum_{s=1}^{n} (x_{is} - x_{js})^2 \right)^{\frac{1}{2}} \tag{1}$$

The purpose of MDS is to find such a matrix X that $d_{ij}(X)$ corresponds δ_{ij}. This assumption can be presented in various forms, including in the least squares MDS model proposed by Kruskal [16]:

$$\sigma^2(X) = \sum_{i=2}^{k} \sum_{j=1}^{i-1} w_{ij} (\delta_{ij} - d_{ij}(X))^2 \tag{2}$$

where w_{ij} is a non-negative weight. For example, many MDS implementations take $w_{ij} = 0$ for the missing differences.

The main assumptions of geometrical models is to meet the following d_{ij} distance conditions:

- Non-negativity
$$d_{ij} > d_{ii} = 0 \text{ (for } i \neq j \text{)} \tag{3}$$
- Symmetry
$$d_{ij} = d_{ji} \tag{4}$$
- Triangular condition
$$d_{ij} \leq d_{ih} + d_{hj} \tag{5}$$

These assumptions have been criticized by Tversky [24], [25], as affecting the empirical observations of similarity. Simultaneously Tversky proposed a different model of similarity, defined by the characteristics of objects. Each object is described by a set of features, and the similarity between objects *a* and *b*, is expressed by the function of common and distinctive features.

Denoted by the A set of features of object *a* and the B set of features of object *b*, the $s(a,b)$ similarity is a function of three arguments, measuring the level that two sets of features fit together (Fig. 1):

- $A \cap B$ – common features *a* and *b*
- $A - B$ – features of *a* not occurring in *b*
- $B - A$ – features of *b* not occurring in *a*

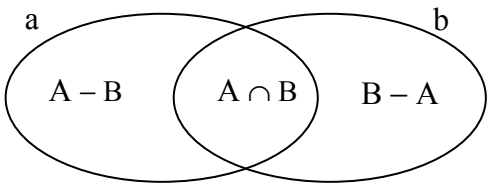


Fig. 1. Graphical representation of sets of features of objects *a* and *b*

Interval similarity scale $S(a,b)$ (*contrast model*), preserving the order of similarity [$S(a,b) > S(c,d)$ if $s(a,b) > s(c,d)$] is expressed as a linear combination of the measures of common and distinctive features:

$$S(a,b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A) \tag{6}$$

where: $\theta, \alpha, \beta \geq 0$, and f is a function representing the contribution of different features of objects in their similarity.

This model does not define a unique index of similarity, but their family, as defined by parameters θ, α, β, thereby allowing the introduction of various relations of similarity between the same objects, such as:

- if $\theta = 1, \alpha = \beta = 0$, then $S(a,b) = f(A \cap B)$
- if $\theta = 0, \alpha = \beta = 1$, then $-S(a,b) = f(A - B) + f(B - A)$

Several hypotheses were defined concerning human perception of similarity in terms of *contrast model*, which then were tested in empirical research [24], [25], where it was confirmed that for man:

I. More important are the common features in determining similarity than in determining difference – focusing attention hypothesis;
II. More important are the features of the compared object (subject) rather than the object with which the comparison is made (reference) - asymmetry hypothesis;
III. More important are the features that are relevant for classification – context hypothesis.

Let us assume that $s(a,b)$ and $d(a,b)$ will be respectively measures of similarity and difference. From the focusing attention hypothesis (I) results that $s(a,b)$ grows along with $f(A \cap B)$ and decreases with the increase of $f(A - B)$ and $f(B - A)$, however $d(a,b)$ decreases with the increase of $f(A \cap B)$ and increases along with the increase of $f(A - B)$ and $f(B - A)$. *Contrast model* weights θ, α, β associated with common and distinctive features will also vary when changing the centre of interest. In the case of evaluating similarities we focus more attention on the similar features, and in the case of a difference we focus more attention on the distinctive features, resulting in weight θ of common features is greater for assessing similarity than for assessing the difference and vice versa.

The asymmetry hypothesis (II) implies that the relation of similarity should not be treated as symmetric (as is the case of geometric models). We cannot assign equivalence to claims such as ”*a* is similar to *b*” and ”*b* is similar to *a*”. Selection of the subject and reference depends largely on the relative importance of objects. We are inclined to choose objects more important as the reference and less important as the subject. For $s(a,b)$, a is the subject, b – the reference. Naturally, we focus attention on the subject, therefore the subject features are more important than features of the reference $(\alpha > \beta)$, and the similarity is more reduced by the distinguishing features of the subject than the reference. For example, a toy train resembles a real train more, because most of the attributes of the toy train is in a real train. On the other hand, a real train is not as similar to the toy, since many of its attributes are not included in the toy [25]. In the *contrast model:* $s(a,b) = s(b,a)$ if and only if $f(A - B) = f(B - A)$ *or* $(\alpha = \beta)$. This means that the symmetry is preserved only when the objects are equally important and their comparison is non-directional, that is, we evaluate the level at which *a* and *b* are similar, and not the level at which *a* is similar to *b* and vice versa.

The context hypothesis (III) tells us, however, that the significance of individual features may vary depending on the considered set of objects and methods of evaluation. An example might be to assess the similarity between two countries bordering close to each other, while having different political systems, such as North Korea and South Korea [24]. Both countries will be judged as more similar to each other in context of European countries or African countries than Asian. Weight of features changes as follows:

- a feature may become more important in some context, if it is the basis for classification in this context
- features that are shared by all objects under consideration do not have a classification value
- when we expand the context, some features may take on the classification value, because they cannot be divided by new objects, so that increases the similarity of objects from their original context
- therefore the similarity of a pair of objects in the original context will be usually smaller than in the extended.

Previous hypotheses were associated with parameters θ, α, β, hypothesis (III) concerns the issue to what degree does *f* change depending on the context of the features.

The *contrast model* also has some gaps, and in many cases the psychological representation is better characterized by structured hierarchical systems. For this reason, *structural models of similarity* [20] were created, which assume that the process of assessing the similarity of a pair of objects must take into account the relations between attributes, and not just attributes. The following example explains this assumption: to model the representation of the ”red square on a blue triangle” attributes of ”red” and ”square” should be combined with the object at the top as well as ”blue” and ”triangle” with the object at the bottom, and adjust the upper and lower object in the ”over” relationship [20]. Thus, the

comparison of objects requires a structural adjustment process, what geometric models lack and those based on features. Therefore, by using them, distinguishing ”red square on a blue triangle” from ”red triangle on a blue square” will be unsuccessful. Capturing these properties requires a structural representation.

According to research, the definition of similarity involves the same process of structural mapping, which is used in inference through analogy [20], [10], [17]. Because of this, structural models of similarity introduce an additional division for common and distinctive attributes of objects. We have two kinds of common attributes [10]:

- MIP (*Match In Place*) is a match between common attributes
- MOP (*Match Out of Place*) is a match between differing attributes.

For example, by comparing a bird with a gray head and red wings with a bird with a gray head and a red tail, the colours of heads are MIP, while the red wings and red tail are MOP. MIP has a greater impact on the similarity than MOP. There are also two kinds of differences between the compared objects [20]:

- agreeable differences – differences between common attributes of objects
- non agreeable differences – differences between attributes that do not match or differences between an attribute in one representation, which does not correspond to any attribute in the other representation.

An example of the *agreeable difference* for a car and motorcycle is the number of wheels, which they possess. However, an example of a *non agreeable difference* for the same objects may be seatbelts, because a motorcycle does not have a device that corresponds to seatbelts. Similar objects tend to have more agreeable differences than differing objects. *Agreeable differences* are easier to determine, they are more important for similarities than *non agreeable differences*.

For completeness of considerations the *transformation models* should be mentioned, which are based on the theory of the Kolomogorov complexity [17], [13]. According to this point of view the measure with which object *a* is similar to *b* is the number of steps in which *a* can be transformed to *b*. Thus the similarity is a function of transformational complexity. Transformation steps may be different in nature and depend mainly on the representation of objects. For sentences of a given language may be, for example, linguistic operations at the level of words, syntaxes and semantics, for structures in the form of a tree – tree transformation operations, etc.. For example XXXOOXO is more similar to OXOOXXX than OXOOXX because OXOOXXX requires only a reflection of XXXOOXO, and OXOOXX requires a reflection plus the removal of X from the right side OXOOXXX [15].

It should be noted that the structural representations, that pose problems for spatial models or feature based models, are easily carried out in the transformation model. Larkey and Markman found some evidence against the transformation model, showing that the number of steps needed to transform colours and shapes of geometric objects is not relevant to human assessment of their similarity [12].

3. Process Characteristics of Data Visualization

Analyzing the data visualization process, you should pay attention to the aspect of the data type and its dimensionality, which directly affects the available forms of presentation. For purposes of visualization, there are three types of data: *nominal*, *ordinal*, *quantitative* [6]. *Nominal* data type is one that can only be equal to or different from other nominal values, *ordinal* data have an additionally defined order, while with *quantitative* data, we can perform arithmetic operations. For example, attributes describing a car, such as manufacturer and model, are nominal data, the segment is of *ordinal* type, and the distance driven is *quantitative* data type. By analyzing dimensionality, we can distinguish the following types of data [23]:

- 1D – linear or sequential data, such as text or program source code (sets and sequences)
- 2D – flat data, such as a floor plan (maps)
- 3D – physical objects, such as molecules, buildings or the human body (shape)
- Temporal – data that include time lines, such as patient records, project data, historical data
- Multidimensional – with more than three variables, like most relational or statistical databases (*case-by-variables*)
- Trees/Hierarchies – each node (except the root) has its unique parent
- Networks/graphs – structures composed of nodes and connections between them.

The object of our interest are mainly multi-dimensional data, since they correspond to the information about the patient’s health condition and disease entities. While the human eye is well

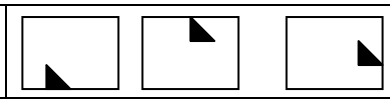
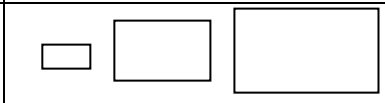
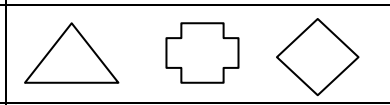
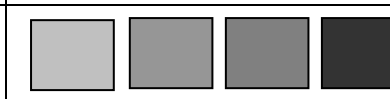
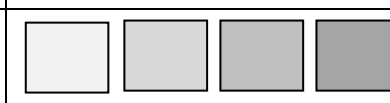
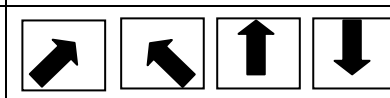
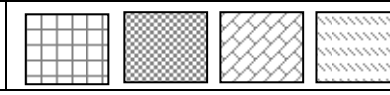
adapted for cases of 1D, 2D and 3D, beyond this limit, we cannot easily map data on graphics structures.

There are many techniques for multidimensional data visualization to deal with the above-mentioned problem. The base for creating visualizations are basic units of information representation called marks recognized by Bertina [4], [5]. Featured symbols are:

- points – denoting the position in space
- lines – representing information of a certain length
- areas – having a length and width (2D)
- surfaces – areas in 3D without thickness
- volume – having a length, width and depth.

In addition, methods of modifying *symbols* were defined called *visual variables* [5]. These include: shape, size, texture, intensity/value, colour, orientation, position. According to Bertin's theory, the human eye is sensitive to these variables, so they are received by the eye effortlessly and automatically (Tab. 1).

Tab. 1. Examples of *image variables*

Position	
Size	
Shape	
Intensity/ Value	
Colour	
Orientation	
Texture	

It was observed that the *visual variables* have a different information transfer efficiency for quantitative data depending on the type of graphical coding. On this basis, the Cleveland & McGill ranking was created, which was later expanded and supplemented by Mackinlay [23], [8], [19]. Mackinlay's ranking (Tab. 2) also includes *nominal* and *ordinal* data types, moreover it expands the image *variables*.

Tab. 2. Mackinlay's ranking of information transmission efficiency of *visual variables* − in order from most to least efficient [19]

Quantative	**Ordinal**	**Nominal**
Position	Position	Position
Length	Intensity/Value	Colour (hue)
Angle	Colour (saturation)	Texture
Slope	Colour (hue)	Connection
Area (Size)	Texture	Containment
Volume	Connection	Intensity/Value
Intensity/Value	Containment	Colour (saturation)
Colour (saturation)	Length	Shape
Colour (hue)	Angle	Length
Texture	Slope	Angle
Connection	Area (Size)	Slope
Containment	Volume	Area (Size)
Shape	Shape	Volume

According to [18], [7] the most significant aspect of visualization is the use of space. It is treated in a particular way in relation to other image attributes – it is the basis, on which other elements are then distributed. So, the empty space of the image is a container with a metric structure that can be described by axis:

- unstructured axis (no axis);
- nominal axis (the region is divided into a sub-region);
- ordinal axis (the order of sub-regions is considerable);
- quantitative axis (region with the metric)

Axes can be linear or radial shape.

Moreover, techniques have been developed that allow increasing the amount of coded information on each axis [7]:

- *Composition* – orthogonal arranging the axis, which allows direct modeling of the relation between data
- *Alignment* – repeating the axis in different places in space, for example, side by side
- *Folding* – continuation of the axis in the dimension perpendicular to it (when there is no place for it)
- *Recursion* – repeated division of space (e.g. reflecting the directory structure)
- *Overloading* – reuse the same space many times (worlds in worlds), based on the fact that the data covers only a portion of space, which allows for the development of the remainder of it.

It should also be mentioned of the importance of the use of symbols and lines to present certain topology [7]. They allow the

representation of relations between objects without geometric constraints (mapping data on the axes of space): *Connections*, that is the indication of relations between objects by drawing links between them, and *Containment*, consisting of representing relations by drawing objects contained within themselves. The form of presentation of the structure may have an impact on its reception by the observer. Bertin distinguished five forms of structural representation of the image (Fig. 2) [5]:

- linear structure – organizes the elements without the use of position
- circular structure – simple in design and allows presenting relations with straight lines
- pattern – the form, in which just the position does not carry any information, but the created pattern can present symmetry or similarity in the structure
- ordered pattern – two-dimensional representation, where one dimension is in order
- stereogram – uses arrangement to present volume and enables observation of 3D patterns.

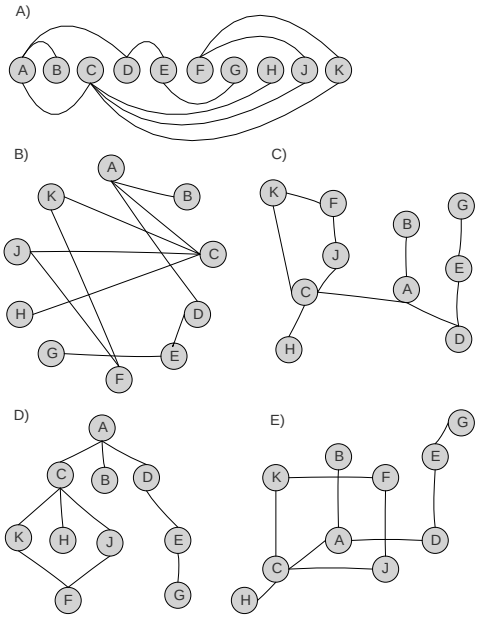


Fig. 2. Forms of structural representation of Bertin's image [5]: A – rectilinear, B – circular, C – pattern, D – orderly pattern, E – stereogram

Among the techniques for multidimensional data visualization, we have a large collection of ready solutions at our disposal, such as: Charts, Treemaps, Scatterplot matrix, Reordable matrix, Parallel coordinates, Glyphs, Spiderweb Chart, Pixel-oriented Technique [23], [3]. Analysis of individual solutions goes beyond the scope of this work, so there will not be further developed of the topic here.

4. Models of Medical Patterns

Now let us consider how the medical data look like that will be visualized and compared. In this paper we will use a simplified model of the description of the patient's condition and the disease entity model proposed by Ameljańczyk [1], [2]. In full version, the model is based on two elements: a set of symptoms and a set of risk factors. Symptoms are all signs of illness identified during the visit to the doctor and as a result of conducted specialized tests, such as high body temperature, swollen glands, coughing, runny nose, fluid in the sinuses, etc. Risk factors are occurrences, which allow to predict the likelihood of the development of the disease entity, for example: obesity, smoking, alcohol abuse, lack of physical activity, etc. To simplify the discussion in the paper we will consider a model limited only to symptoms.

If we assume that the set $S \subset \mathcal{N}$ is a set of numbers of all the symptoms, which can describe the disease entity, then the disease entity model $m \in M = \{1,...,M\}$ can be formally written as:

$$M(m) = \left(S^m, C^m\right), \tag{7}$$

where:

S^m – set of symptoms numbers of disease $m \in M$

C^m – set of disease value ranges of symptoms of disease $m \in M$.

Additionally, S^m and C^m:

$$S^m = \left\{s_1^m,...,s_k^m,...,s_{K(m)}^m\right\} \subset S, m \in M \tag{8}$$

$$C = \left\{C_1^m,...,C_k^m,...,C_{K(m)}^m\right\} \; m \in M \tag{9}$$

$K(m)$ – number of symptoms of disease entity $m \in M$.

$C_k^m = \left\lfloor \underline{c}_k^m, \overline{c}_k^m \right\rfloor$ – disease value range of symptom *k* in disease $m \in M$.

The model of the patient's condition $x \in X$, built based on the disease entity model, will be presented in the form of:

$$P(x) = \left(S_o(x), W(x)\right) , \tag{10}$$

where: set $S_o(x) \subset S$ is a set of disease symptoms numbers occurring in a patient, and $W(x)$ is a set of levels of severity of different symptoms:

$$S_o(x) = \left\{s \in S \middle| w(x,s) > 0\right\} \tag{11}$$

with $w(x,s) \in W(x)$ – level of severity of symptom $s \in S$.

It should be noted that if $s = s_k^m$ and $w(x,s) \in C_k^m$, then the symptom severity is in the range of the disease values for the disease entity $m \in M$.

5. Medical Data Visualization Space

Changing the disease entity model and the model of the patient's health condition to a graphical representation will be started by introducing the *Information Visualization Design Space* [6]. We will use a narrowed down description proposed by Mackinlay, i.e. *Space Visualization*, which is based on:

- *Symbols:* Point, Line, Area, Surface, Volume
- *Visual variables:* Color, Size, Shape, Intensity, Orientation, Texture, Connection, Inclusion, Position (X, Y, Z).

For comparison reasons of individual visualizations, Mackinlay presented principles of visualization in the form of a table, which, adapted to our needs, will contain columns as in Tab. 3. Their explanation is presented in Tab. 4.

Tab. 3. Visualization description in the form of a table [6]

Variable	D	F	D’	R	X	Y	Z

Tab. 4. Explanation of symbols for the tabular description of the visualization

Symbol	Definition
Variable	Name of the represented information
D	Data type: N (Nominal), O (Ordinal), Q (Quantative)
F	Function re-coding data, e.g.: f (unspecified), > (filter), s (sorting), mds (MDS)
D’	Data type re-coded
R	Visual variable: C (Color), S (Size), F (Shape), (V) Intensity, O (Orientation), T (Texture), -- (Connection), [] (Inclusion)
X,Y,Z	Position in space represented by a symbol: P (Point), L (Line), S (Surface), A (Area), V (Volume)

For example, information that will be presented in our visualization derive from a disease entity model and a model of a patient's health condition. Visualization attributes of the disease entity (limited only to symptoms and their standard disease values), written in the form of a table could look like Tab. 5.

Tab. 5. Example of the visualization description in a tabular form for the disease entity

Variable	D	F	D’	R	X	Y	Z
Symptom	N				A	A	
Standard disease symptom value	Q	f	O	C			

Tab. 5 presents only one of the possible variants of visualization, in which the symptoms as a *nominal* (N) data type are mapped to areas (A) in space (X,Y), while the standard disease symptom value (quantitative type) (Q) is transformable using the function *f* to the *ordinal* (O) set of data and mapped to a color (C).

The description in the form presented above can be used only to compare the properties of different visualizations. It is certainly insufficient to construct a target image. There are a few missing aspects, among others there are no precisely defined methods of arranging symbols or a method of using information coding techniques. It is therefore necessary to supplement it with transformation rules that will transform an object into a particular image. A full set of rules generating a unique image on the basis of object attributes will be called the *visualization model*.

For the purpose of further consideration, let us assume that any object *o* belonging to the universe $o \in U$, is represented as a set of attributes $o = \{c_1, c_2, ...\}$ and denoted by *g* of any image belonging to the set of all possible images $g \in G$ to generate. Our projection, converting the object to an image based on the visualization model, can be written in the form of $v : U \to G$, i.e. $v(o) = g$.

6. Study of Visualization Model

Let us denote by $V = \{v_1, v_2, ...\}$, the set of all possible *visualization models* of disease entities and patient health condition, where $v \in V$ is the *visualization model* defined as in the shown above description. With $v(m)$ we will be denoting the application of model *v* for the disease entity $m \in M$, which is a specific graphical representation (image) of a disease

entity *m*. For simplification reasons we will assume $p = P(x)$ as patient *x*'s health condition. Its graphical representation generated by the visualization model *v* will be indicated $v(p)$. We assumed that *m* and *p* are objects, for which the same *visualization model v* can be applied. This can be accepted since both types of objects are made up of symptoms and respectively – the standard disease value and the current level of symptom intensification.

Our task is to find an optimal *visualization model* $v^* \in V$ that maximizes the similarity assessment of the disease entity and the patient's health condition, if the medical condition corresponds to the given unit. This means that when a physician compares the graphical representation of the patient's health condition with examples of graphical representations of various disease entities, then the most similar disease entity for him/her will be the one that the patient is suffering from.

Previously discussed similarity models can be used as a basis for evaluation and formulation of constraints for verified solutions. In accordance to the *contrast model*, we can define a similarity scale projecting a natural (perceived by the observer) order of compared visualizations of disease entities and health condition in the form of $s(a,b) = s(v(p),v(m))$. The value v^* that is search by us must meet condition:

$$s\left(v^*(p),v^*(m)\right) \geq s\left(v\ (p),v\ (m)\right) \tag{12}$$

for every $p = P(x)$ and for every $m \in M$, when patient *x* is ill with disease entity *m*, where $v \in V_m \setminus \{v^*\}$.

Another condition, which we will defined is:

$$s\left(v^*(p),v^*(m_1)\right) > s\left(v^*(p),v^*(m_2)\right) \tag{13}$$

for every $p = P(x)$ and for every $m_1, m_2 \in M$, when patient *x* is ill with disease entity m_1 and is not ill with m_2.

The final condition written as follows:

$$s\left(v^*(p_1),v^*(m)\right) > s\left(v^*(p_2),v^*(m)\right) \tag{14}$$

corresponds to the case, in which $p_1 = P(x_1)$ and $p_2 = P(x_2)$ represent the health condition of patients x_1 , x_2 ill with the same disease entity $m \in M$, where patient x_1 has more symptoms corresponding to disease entity *m* than patient x_2, that is, the following inequality occurs between cardinalities of sets of common symptoms of health condition and disease entity:

$$card\left(S_o(x_1) \cap S^m\right) > card\left(S_o(x_2) \cap S^m\right) \tag{15}$$

On the basis of (12), among all visualization models of the disease entity with the application of visualization model *v** for each case of the patient's health condition and disease entity (which corresponds to this case), the biggest similarity is observed. From (13) results that within the visualization model *v**, the biggest similarity will always concern the disease entity similarity, which corresponds to the case of the patient's health condition. However, the last condition (14) is fulfilled, if with the increasing number of symptoms corresponding to the disease entity, the similarity increases of the patient's health condition and the entity.

Keeping in mind the form of the linear combination of *contrast model* (4), and the hypotheses of similarity (I, II, III), we can consider the task of finding the optimal visualization model. Let us assume that:

$$attr(g) = \{a_1, a_2, ...\} \tag{15}$$

where $g \in G$ and $n \in N$, is the operator determining the set of image attributes identified by the observer. Then for weights θ, α, β, scale f and defined $G^p = attr(v(p))$ and $G^m = attr(v(m))$ the model has a form of:

$$\begin{aligned} S\left(G^p, G^m\right) &= \theta f(G^p \cap G^m) \\ &- \alpha f(G^p - G^m) \\ &- \beta f(G^m - G^p) \end{aligned} \tag{16}$$

Therefore increasing the observed similarity will consist of:

- $\theta \rightarrow \max$
- $\alpha, \beta \rightarrow \min$
- $f\left(G^p \cap G^m\right) \rightarrow \max$
- $f\left(G^p - G^m\right) \rightarrow \min$
- $f\left(G^p \cap G^m\right) \rightarrow \min$

It should be noted that the set of visualized object attributes, in our case – a set of symptoms, does not translate easily to the attributes of the visualization itself. The visualization may contain attributes that are an indirect result of its structure. According to the creators of *Gestalt psychology*, human image perception should be treated as a whole, without decomposing into smaller entities [9]. According to this theory, only the global relation between all elements determines the main aspects of perception, e.g. perception of a circle created out of single points, which are equally distant from the center. Therefore, for the visualization of the disease entity size of the sets of the disease entity

$$card\left(G^p \cap G^m\right) \tag{17}$$

and

$$card\left(S_o(x) \cap S^m\right) \tag{18}$$

for $m \in M, p = P(x)$, they do not have to be equal. Thus, in order to try to calculate the similarity, we have to complete the task of determining all the characteristics that are recognized by the observer, i.e. constructing the operator *attr*.

As previously discussed the visualization model is defined as a set of rules that generate the image for a certain object. By Analyzing the various hypotheses of similarity, we consider their impact on the optimal construction of such rules. On the basis of (I), we can assume that visualization models will be preferred that display common attributes. This is because in our case, the task put forward to the observer is to assess the similarity and not assessing the difference. The relation directionality resulting from hypothesis (II) has already been implicitly introduced in our consideration by defining the task as an "assessment of the similarity of the patient's health condition with the disease entity." In this situation, the subject is the patient's health condition and the disease entity is the reference. Thus, a greater impact on the assessment of similarity are the distinguishing attributes of the patient's health condition than the attributes distinguishing the disease entity, which is another indication affecting the construction of the visualization model.

Using hypothesis (III) also becomes a source of guidance. First of all attributes should be preferred in the presentation that are relevant to the classification; secondly, an important role in the process may have the presentation of each image to the observer. For example, we can imagine that the presentation of the patient's health condition in the context of a few different images of disease entities will change the assessment of similarity in relation to the presentation in the context of a single entity.

7. Conclusion

This paper presents preliminary results in the scope of assessing the role of data visualization in determining the similarity of medical patterns. Described similarity models can serve as a guide to the evaluation of visualization methods. The emphasis has been on the *contrast model*, however, during further research other models should not be forgotten. For example, it may be interesting to apply the MDS model in the first stage of selecting disease entities, with which the patient's health condition will be matched with. Visualizing all disease entities at the same time seems impossible, therefore, reduction of this set at the initial stage should be applied. For this purpose we could use another type of image created by MDS. The graphics would present the patient's health condition and disease entity as a single point on the plane. Therefore, the use of MDS requires knowledge of the value of difference/similarity δ_{ij} between individual objects, for our case they could be calculated on the basis of the number of corresponding symptoms in individual objects. Such a proposed visualization would present in the immediate neighbourhood of the point representing the patient's health condition points representing disease entities having the most common symptoms with it. The second stage would be to present the observer the visualization of the patient's health condition in context of a few selected disease entities from the first stage – many simultaneously, or sequentially, with each individually.

The main issues for further research include the development of the subject of visualization space and the creation of description principles of a full visualization model. Another area is to formulate a complete optimization problem, which solution would be the best visualization model (in accordance with pre-defined understanding of this concept).

Closely associated with this is also the construction of a method of models assessment. This is a nontrivial task, if we assume to try to at least partially automate it. It is related to the issue of identifying attributes of objects, i.e. building the *attr* operator. It will probably be useful here to refer to publications dealing with the human perception of the image, such as the work of Colin Ware [26]. At the same time, because of ready algorithms such as: SME (*Structural Mapping Engine*) [20], SIAM (*Similarity as Interactive Activation and Mapping*) [10], CAB (*Connectionist Analogy Builder*) [10], the use of *structural* and *transformational models* can be an area of additional research opportunities, particularly in the scope of visualization model assessment.

In conclusion, the presented work attempts to define the basis for further research of medical data visualization. The expected result is to find a visualization model that allows for the creation of a new tool to support physicians in diagnosis and treatment, thus contributing to the elimination of some popular medical malpractice.

8. Bibliography

[1] A. Ameljańczyk, „Analiza wpływu przyjętej koncepcji modelowania systemu wspomagania decyzji medycznych na sposób generowania ścieżek klinicznych”, *Biuletyn Instytutu Systemów Informatycznych*, Nr 4 (2009).

[2] A. Ameljańczyk, „Wielokryterialne mechanizmy wspomagania podejmowania decyzji klinicznych w modelu repozytorium w oparciu o wzorce”, *Biuletyn Instytutu Systemów Informatycznych*, Nr 5 (2010).

[3] M. Ankerst, ”Visual Data Mining with Pixel-oriented Visualization Techniques”, *ACM SIGKDD Workshop on Visual Data Mining*, San Francisco, CA, 2001.

[4] J. Bertin, *Graphics and Graphic Information-Processing,* Walter de Gruyter, Berlin, 1981.

[5] J. Bertin, *Semiology of Graphics: Diagrams, Networks, Maps*, The University of Wisconsin Press, Wisconsin, 1983.

[6] S. Card, J. Mackinlay, ”The Structure of the Information Visualization Design Space”, *INFOVIS '97*, IEEE Computer Society Washington, DC, USA, 1997.

[7] S. Card, J. Mackinlay, B. Shneiderman, *Readings in information visualization: using vision to think*, Morgan Kaufmann Publishers, San Francisco, 1999.

[8] W. Cleveland, R. McGill, ”Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods”, *Journal of the American Statistical Association*, 79, 531–554 (1984).

[9] E. Goldmeier, ”Similarity in Visually Perceived Forms”, *International Universities Press, Inc.*, 1972.

[10] R. Goldstone, ”Similarity, Interactive Activation, and Mapping”, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 20, No. 1, 3–28 (1994).

[11] P. Groenen, M. van de Velden, ”Multidimensional Scaling”, *Econometric Institute Report*, EI 2004-15, April (2004)

[12] J. Groopman, *Jak myśli lekarz*, Wydawnictwo Dolnośląskie, Wrocław, 2009.

[13] U. Hahn, N. Chater, L. Richardson, ”Similarity as transformation”, *Cognition*, 87, 1–32, Elsevier, 2003.

[14] D. Hintzmann, ”Schema abstraction in a multiple-trace memory model”, *Psychological Review*, 93, 411–428 (1986).

[15] S. Imai, ”Pattern similarity and cognitive transformations”, *Acta Psychologica*, 41, 433–447 (1997).

[16] J. Kruskal, ”Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis”, *Psychometrika*, 29, 1–27 (1964)

[17] L. Larkey, A. Markman, ”Processes of Similarity Judgment”, *Cognitive Science*, 29, 1061–1076, Cognitive Science Society, Inc., 2005.

[18] A. MacEachren, *How Maps Work*, The Guilford Press, New York, 1995.

[19] J. Mackinlay, ”Automating the Design of Graphical Presentations of Relational Information”, *ACM Transactions on Graphics* , Vol. 5, Issue 2, New York, USA, April, 1986.

[20] A. Markman, D. Gentner, ”Structural Alignment during Similarity Comparisons”, *Cognitive Psychology*, 25, 431–467, Academic Press, 1993.

[21] D. Medin, R. Goldstone, A. Markman, ”Comparison and choice: Relations between similarity processes and decision processes”, *Psychonomics Bulletin and Review*, 2, 1–19 (1995).

[22] R. Nosofsky, ”Attention, similarity and the identification-categorization relationship”, *Journal of Experimental Psychology*, 115, 39–57 (1986).

[23] H. Siirtola, ”Interactive Visualization of Multidimensional Data”, *Dissertations in Interactive Technology*, Vol. 7, 2007.

[24] A. Tversky, ”Features of Similarity”, *Psychological Review*, Vol. 84, Number 4, 327–352, American Psychological Association , July, 1977.

[25] A. Tversky, I. Gati, ”Studies of Similarity”, in: *Cognition and Categorization*, 79–98, E. Rosch & B. B. Lloyd (Eds.), Hillsdale, 1978.

[26] C. Ware, *Information Visualization: Perception for Design, 2nd Edition*, Morgan Kaufmann Publishers, 2004.

Wizualizacja danych w określaniu podobieństwa wzorców medycznych

T. RZEŹNICZAK

W artykule przedstawiono koncepcję wykorzystania teorii podobieństwa w rozpoznawaniu wzorców medycznych. Celem prowadzonych prac jest skonstruowanie postaci graficznej wzorca jednostki chorobowej oraz stanu zdrowia pacjenta, w taki sposób, aby wykorzystać naturalne zdolności percepcyjne człowieka do identyfikacji podobieństwa między nimi. Dzięki takiemu podejściu, reprezentacja wzorców medycznych może zostać zastosowana do wsparcia procesu diagnozowania jednostek chorobowych.

Słowa kluczowe: wizualizacja danych, modele podobieństwa, relacja podobieństwa, diagnostyka medyczna.

Concept of Usage of Bayesian Networks in Clinical Decision Support Module

M. STRAWA
marcin.strawa@gmail.com

Institute of Computer and Information Systems
Faculty of Cybernetics, Military University of Technology
Kaliskiego Str. 2, 00-908 Warsaw, Poland

Concept of decision support module utilizing a repository of clinical pathways has been presented in this paper: the definition of Bayesian networks and its major concepts, description of chosen inference algorithm and an example of diagnosis.

Keywords: Bayesian networks, belief networks, clinical decision support system.

1. Introduction

Bayes' theorem expresses the conditional probability of hypothesis H (given evidence E) in terms of the prior probability of H, the prior probability of E, and the conditional probability of E given H.

$$P(H \mid E) = \frac{P(H)P(E \mid H)}{P(E)} \tag{1}$$

Bayesian network is a tool that is based on this theorem. In this paper, the concept of usage of Bayesian networks in a clinical decision support module is presented. The purpose of the module is to cooperate with clinical pathways repository and taking decisions, which are located in decision nodes of appropriate pathways, as well as the selection of the proper pathway to follow. It is illustrated in an example, where preliminary diagnosis is taken and basing on this, the clinical pathway is chosen. Also, the algorithm is presented for making decisions in a single decision node of the pathway. The disease chosen for the example is chronic myeloid leukemia.

The paper begins with the definition of Bayesian networks and its major concepts. Next, the short description of chronic myeloid leukemia takes place.

In the following section, based on the description of the diagnostic process, the simple clinical pathway has been constructed as an example. It is a basis for the illustration of the reasoning procedure, which is shown further. The next section shows the reasoning algorithm and its execution for the exemplary Bayesian network.

2. Bayesian Networks

The Bayesian network (other name: *belief network*) is a probabilistic graphical model representing a set of random variables and its conditional dependencies as a acyclic directed graph. Vertices of a Bayesian network represent all attributes defined in the problem's domain, while edges can be interpreted as a representation of the direct causal dependency between them.

There are a few formal definitions of the Bayesian network. For all the definitions given below let us assume that $G = (V, E)$ is an acyclic directed graph, and $X = (X_v)_{v \in V}$ is a set of random variables indexed by V.

1. X is Bayesian network with respect to G, if its joint probability distribution can be defined as a product of conditional probabilities of all the nodes:
$$P(x) = \prod_{v \in V} P\left(x_v \middle| x_{pa}(v)\right) \tag{2}$$
where $pa(v)$ is a set of all parents of node v.
2. X is Bayesian network with respect to G, if it satisfies the *local Markov property*: each node is conditionally independent of its non-descendants given values of its parent nodes:
$$P\left(X_v = x_v \middle| \forall j \in V - \{v\} - S_v : X_j = x_j\right) = P\left(X_v = x_v \middle| \forall j \in U_v : X_j = x_j\right) \tag{3}$$

where:

S_v – set of numbers of all (direct or indirect) descendants of node v

U_v – set of numbers of all parents of node v

The main advantage of Bayesian networks (having a proper structure) is the ability of

representing indirectly the joint probability distribution of all variables in an efficient way. To represent such a distribution with the Bayesian network, for each node v it is required to know conditional probabilities of its values, given the values of its parent nodes. It is sufficient then to store Vk^{u+1} probability values (where: $k = \max_{v \in V} |X_v|$, $u = \max_{v \in V} |U_v|$), while the direct representation of the joint probability distribution would require to store the following number of probability values: $\prod_{v=1}^{V} |X_v| \leq k^n$.

Example[1]**:**

House alarm systems react to burglaries as well as earthquakes. Neighbours Mary and John are agreed to call the owner when they hear the alarm. John always calls, but sometimes takes the ringing phone for an alarm signal and calls then, too. Mary likes loud music and then sometimes misses the alarm. Given the evidence who has or has not called we want to estimate the probability of a burglary. The Bayesian network for this example is presented below:

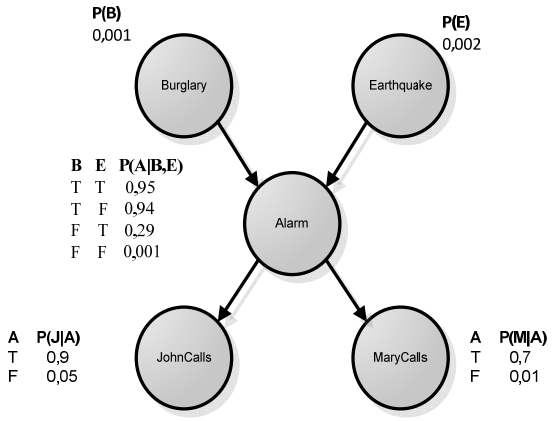


Fig. 1. Example Bayesian network

Let us say we want to calculate the probability of the alarm when there was no burglary and earthquake, given that both John and Mary called:

$$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) =$$
$$= P(J \mid A)P(M \mid A)P(A \mid \neg B \wedge \neg E)P(\neg B)P(\neg E) =$$
$$= 0.90 \cdot 0.70 \cdot 0.001 \cdot 0.999 \cdot 0.998 = 0.00062$$

Knowledge of joint probability distribution for all variables makes it possible to carry out probabilistic reasoning for values of every combination of variables, given the values of other variables.

Let $V_0(x_q)$ – set of numbers of variables having known values;

$V_h(x_q) - V - V_0(x_q)$ – set of numbers of variables, which values are not known for some example $x_q \in X$. We look for probability distribution of variables numbered by $V_h(x_q)$, having given values of variables numbered by $V_0(x_q)$. To calculate it, the following formula can be used:

$$P\left(\forall i \in V_h(x_q): X_i = x_i \middle| \forall i \in V_0(x_q): X_i = x_i\right) =$$
$$= \frac{P(\forall i \in V: X_i = x_i)}{P(\forall i \in V_0(x_q): X_i = x_i)}$$

for all values of $x_i \in X_i$ if $i \in V_h(x_q)$ and for known values of $x_i = x_q$ if $i \in V_0(x_q)$.

The answer for every query can be obtained by calculating, with the usage of the network, joint probability distribution and applying it for subsequent calculations. Unfortunately, this approach means giving up one of the main advantages of representing joint probability distribution as a Bayesian network – efficiency. Due to this fact, other algorithms are used for answering such queries. In general, reasoning in Bayesian networks is NP-hard, so approximation algorithms are mainly utilized in problem solving. There is also one type of Bayesian network for which the reasoning problem is much simpler, so that effective exact algorithms can be applied. These are networks, where only one undirected path exists between any pair of nodes.

Simple examples of four reasoning patterns, for which Bayesian networks can be utilized are shown on graph 2. *E* stands for observable attribute (evidence variable), *Q* – for attribute, which is a subject of a query (question variable).

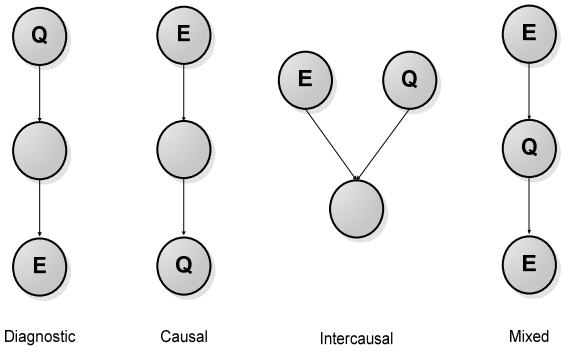


Fig. 2. Reasoning patterns that can be handled by a Bayesian network

[1] Example taken from [2]

3. Chronic Myeloid Leukemia

Leukemia[2] is a malicious tumor of hematopoietic cells being formed as a consequence of systemic, scattered and autonomous growth of one leucocyte clone and the spreading of cancer-altered, immature blast cells from bone marrow into the blood.

Main forms of leukemia can be divided into four categories:

- acute myeloid leukemia (AML)
- chronic myeloid leukemia (CML)
- acute lymphocytic leukemia (ALL)
- chronic lymphocytic leukemia (CLL)

In contrary to acute leukemia, progress of chronic myeloid leukemia (CML) is long lasting and relatively slow.

Despite CML it is quite a frequent category of leukemia, its occurrence is rare, taking into account global population. Most patients are adults. Children are only 2–4% of the cases.

Chronic myeloid leukemia is caused by changes in the genetic code of some cells in the bone marrow. In these cells, a part of chromosome 9 becomes a place of part of chromosome 22 – a process called translocation. Abnormal chromosome called the Philadelphia chromosome is formed. Abnormal chromosome stimulates the overproduction of white blood cells in the bone marrow.

Chronic myeloid leukemia generally proceeds in three phases. Most patients are diagnosed in the initial phase called chronic. Over time it transforms into the acceleration phase – the disease accelerates, and finally into the blastic phase – the most malicious and of a course similar to acute leukemia.

Chronic phase is a first phase of disease and lasts much longer than others. There is a larger number of white blood cells in blood and bone marrow, but most of them are mature cells that function properly. Most patients (80%) remain in the stable phase for at least 5 years. Symptoms of the chronic phase of CML depend on what kind of white blood cells are present in the blood of a given patient. Typically, the symptoms are scarce, and the disease is diagnosed by routine blood tests

Symptoms may include:

- fatigue
- headache
- pain or feeling of fullness in the left mid-abdomen (caused by an enlarged spleen).

Acceleration phase. At this stage there is an increasing number of immature cells (blasts) in the blood, bone marrow, liver and spleen. Blasts cannot fight infections like normal white blood cells. In the past, the length of acceleration was usually one to six months before progressing to the blastic phase. Depending on the treatment, this phase can be extended to more than 1 year.

Signs of accelerated phase are more intensive and include:

- fever
- night sweats
- weight loss
- pale skin, easy fatigue, shortness of breath (deficiency of red blood cells, or anemia).

Blastic phase. In this phase comes the rapid progression of the disease and the creation of huge numbers of malignant cells in the blood. The result is an increasing number of blasts from the marrow and the displacement of normal blood cells in the blood – red cells, white blood cells and thrombocytes. Patients often report problems with infections, easy bruising and bleeding. The course of the disease resembles acute myeloid leukemia, or in rare cases, acute lymphoblastic leukemia.

In order to recognize chronic myeloid leukemia and to assess the progress of the disease, a puncture is carried out (bone marrow picking).

4. Clinical Pathway

Based on the description above, a sample fragment of the clinical pathway for CML, enclosing the diagnosis stage, can be constructed. It will be useful further in this paper as an input to the decision support module utilizing Bayesian networks.

[2] Description taken from [5] and [4]

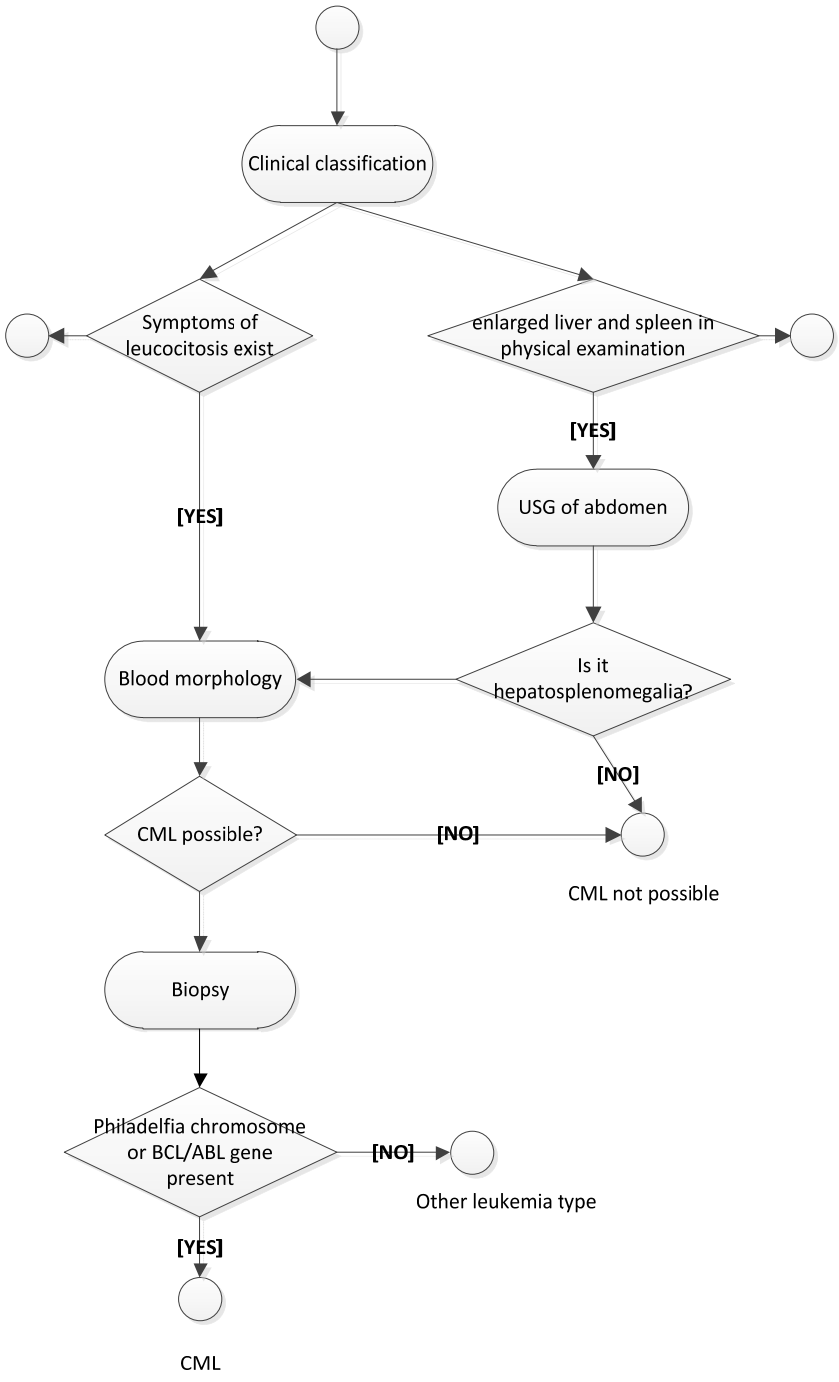


Fig. 3. Sample fragment of the clinical pathway for CML

5. Example of diagnosis with the usage of bayesian network

In order to show a mechanism of reasoning, sample bayesian network for CML has been constructed. Probability values have been taken arbitrary, only to illustrate the example. In working system they must be determined with help of domain experts as well as by network learning (sample methods are presented in further part of this paper).

The purpose of the constructed network is to conduct diagnostic reasoning. For evidence variables stand observable disease symptoms (through patient interview or examinations results), for question variables – diseases. Continuous variables have been transformed into discrete variables by dividing their set of values into ranges bounded with values having medical significance (limits of various norms for a healthy adult person, typical values for analyzed diseases, etc.).

Algorithm. For every iteration of the reasoning process, the disease having highest occurrence probability, in context of known symptoms, is searched. Then the clinical pathway, which is most suitable for found disease, is selected. If the probability value is not equal 1 or a defined level for proper diagnosis, the decision will be taken to make examinations defined on a selected pathway required to verify the current diagnosis. Results of examinations extend the set of evidence variables and a new iteration takes place. Results of examinations defined on the clinical pathway, which results are known, are marked as done to avoid multiple executions. There are two stop conditions:

1. diagnosis has been found, there is no other one with a higher possibility value,
2. all examinations defined on executed pathways have been done.

The reasoning method applied in each algorithm iteration. As it was stated before, the reasoning problem in Bayesian networks is NP-hard and accurate, effective algorithms exist only for networks having a polytree structure. The network constructed for the analyzed diagnostic problem does not have a mentioned structure, as there can be found at least one pair of nodes having more than one undirected path between them. Usage of approximation algorithms is required then.

Three classes of reasoning algorithms for Bayesian networks are known:

- Clustering methods – the network is transformed into probabilistically equivalent (but topologically different) polytrees by merging nodes. Then known accurate algorithms can be applied
- Conditioning methods – variables in networks are substituted with particular values. Every possible substitution is evaluated
- Stochastic simulation (Monte – Carlo) – big number of samples (networks with defined values of attributes) is generated, for which conditional probabilities in nodes are consistent with the ones in the analyzed network. Distribution of results is an approximation of exact evaluation.

Interesting effectiveness comparison of a few most popular algorithms can be found in [6]. For the analyzed example Monte-Carlo-class algorithm will be used. Its name is *Likelihood Weighting*.

Every iteration of simulation using this algorithm looks like the following:

1. generate values of variables for all root nodes with probability distribution defined in nodes,
2. for each following node:
   a. if the node is not an evidence variable: generate the variable's value according to its conditional probabilities table, assuming known values of conditions,
   b. if the node is an evidence variable: find the probability value in its conditional probabilities table assuming the known value of the observed variable and known values of conditions. The found probability will be the weight of whole simulation step.

After finishing the simulation step, the probability of reaching some value by the question variable, under the condition of evidence variables, is known.

The simulation result is the quotient of the sum of probabilities of interesting events' occurrences by the sum of all probabilities attained in simulation steps.

```
function LikelihoodWeighting(X, e, n, N)
returns estimation P(X|e)
local variables: W, vector of weights of values
in X

for j = 1 to N do
   x, w := WeightedSample(n, e)
   W[x] := W[x] + w where x is value of X in x
return Normalize(W[X])
```

```
function WeightedSample(n, e) returns event
and weight

x := n-element event;
w := 1
for i = 1 to n do
   if X_i has value x_i in e
      then w := w × P(X_i = x_i | Parents(X_i))
      else x_i := random value with P(X_i |
Parents(X_i))
return x, w
```

Bayesian network for chronic myeloid leukemia. Nodes represent chosen symptoms and causes of the disease. The node representing leukemia is node **CML**. There is also a few other diseases present, which can be potential causes of the symptoms.

Inference for the exemplary network. Let us provide patient complains about seeing disorders, night sweats and during the physical examination where the enlarged spleen was noted. So the specified symptoms are evidence variables in our exemplary network. Node CML represents the question variable.

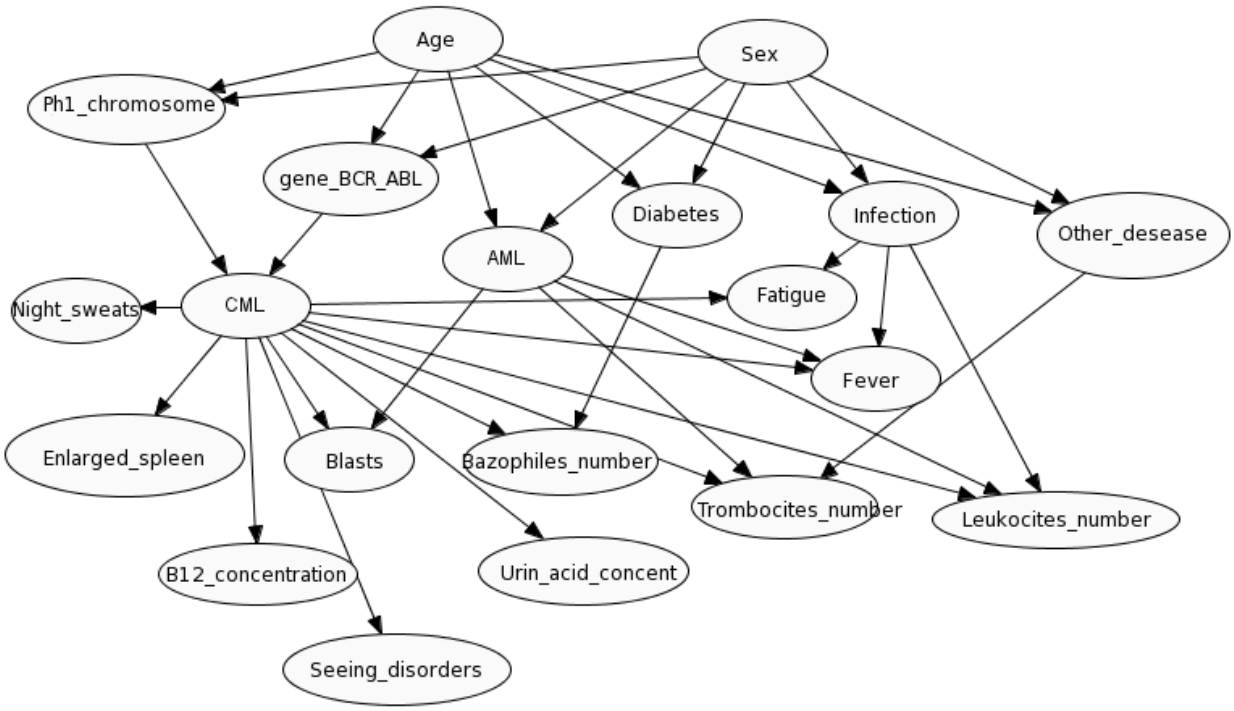


Fig. 4. Sample Bayesian network for chronic myeloid leukemia

In one iteration of the likelihood weighting algorithm, nodes (due to limited space, proceeding for only the chosen ones is presented) will take on the values as the following:

1. Take the weight of the iteration $w = 1$.
2. For node **Age** draw a value according to its probability table. Let us provide value is $Age = 40 - 60$.
3. For node **Sex** draw a value according to its probability table. Let us provide that the value is $Sex = F$.
   For node **Ph1_chromosome** draw a value according to its conditional probability table. For values of predecessors: $Age = 40 - 60$ and $Sex = F$ probability values are: 0.000002, that *Ph*1_ *chromosome* = *True* and 0.999998, that *Ph*1_ *chromosome* = *False*. Generation takes place according to these values. Let us provide that the value drawn is *Ph1_ chromosome = False.*
4. For node **gene_BCR_ABL** draw a value according to its conditional probability table. For values of predecessors: $Age = 40–60$ and Sex = F probability values are: 0.000012, that *gene_BCR_ABL = True* and 0.999988, that *gene_BCR_ABL = False*. Generation takes place according to these values. Let us provide the value drawn *gene_BCR_ABL = True.*
5. For node **CML** draw a value according to its conditional probability table. For values of predecessors: *ph1_chromosoe = False* and *gene_BCR_ABL = True* probability values are 0.99999, that *CML = True* and 0.00001, that *CML = False*. Let us provide the value drawn *CML = True.*
6. **Night sweats** node is an evidence variable as it contains symptoms found during examination. We know that its value equals *True*, and the predecessor's value: *CML = True*. So the weight of iteration w must by modified according to node's conditional probabilities table. $w := w^* P(Night\_sweats = T \mid CML = T) = w^* 0{,}1 = 0{,}1$.
7. **Enlarged_spleen** node is an evidence variable. We know that its value equals *True*, and predecessor's value: *CML = True*. So the weight of iteration w must by modified according to node's conditional probabilities table. $w := w^* P(Enlarged\_spleen = T \mid CML = T) = 0{,}1 * 0.35 = 0.035$.
8. **Seeing_disorders** node is an evidence variable. We know that its value equals
9. *True*, and the predecessor value: *CML = True*. So the weight of iteration w must by modified according to node's conditional probabilities table. $w := w^* P(Seeing\_disorders = T \mid CML = T) = 0.035 * 0.1 = 0.0035$.
10. Return set of all nodes' values together with weight of iteration $w = 0.035$.

After conducting the necessary number of simulations, following the steps above, the probabilities for interesting nodes must be determined. In the analyzed case the interesting nodes are: CML, AML, diabetes, infection, other_disease and more specific – probability that their value equals *True*. Thus, for each one the following value must be calculated:

$$P(X = True \setminus e) = \frac{\sum_{i \in W : X = True} W_i}{\sum_{i=1}^{N} W_i},$$

where:
X – interesting node,
N – number of simulation steps,
e – observable attributes,
W – vector of simulation results.

Let us provide that the calculated probability values are as the following:

- $P(CML = T) = 0.2$
- $P(AML = T) = 0.01$
- $P(Diabetes = T) = 0.007$
- $P(Infection = T) = 0.08$
- $P(Other\_disease = T) = 0.13$

So the decision on the diagnosis cannot be taken, but the node having the highest probability value is CML. The pathway for this disease must be then proposed. As the first examination showed a suspicion of spleen enlargement, the system will advise USG of the abdomen in this decision node. Next, after the results of new examinations are presented in the network as a evidence variables, the reasoning process will be repeated.

The major problem in the application of the presented method is the necessity of performing a big number of simulations (which means a large amount of time) to obtain precise probability values for the least probable events. The time required to reach a particular precision level is reversely proportional to the probability of the event.

6. Learning Bayesian Networks Basing on Examples

Although it was provided that the network would be constructed using knowledge of the domain experts, the ability to automatically construct one may significantly increase its usability. It can be achieved using methods of Bayesian networks learning with the usage of training data.

There are two criteria by which we can group problems of learning:

- knowledge of the network's structure or the lack of it
- all or only part of the attributes are observable in the training data.

For the proposed reasoning module, we have a problem where the network's structure is well defined but not all values of the attributes for the training data are known. The problem can be transformed to the calculation of conditional probability tables for the network with a defined structure and some training set T. The goal of learning is to find a hypothesis h, which is most consistent with training data. It means maximization of probability P(T|h). Descriptions of algorithms used to achieve this goal can be found, for example, in [1], [2], [8].

Many researches are currently made in this area. Interesting methods can be found i.e. in [9].

7. Summary

The concept of the decision support module utilizing the repository of clinical pathways has been presented in this paper. The utility used in this module is the Bayesian network. It has already been successfully applied in supporting medical diagnostic processes in the country, as well as abroad. Examples of these systems are HEPAR and MUNIN. The concept of graphical network presentation makes it easier to construct one in cooperation with domain experts, because it helps to understand causal dependencies between variables.

The problem of inference in Bayesian networks is NP-hard, but the number of effective algorithms producing approximate results of good quality was invented. One of them is, described here, Likelihood Weighting algorithm, based on the Monte-Carlo approach. Its major weakness is the precision of computing probability values of slightly probable events. However, modifications exist, which allow reducing this problem. There are also researches on effective reasoning methods.

The other developing domains are algorithms of learning Bayesian networks, which are very useful for constructing networks with the usage of training data.

Implementation of the described module would allow performing few experiments regarding effectiveness and performance of various learning and inference algorithms, in cooperation with various clinical pathways. It might lead to the formulation of a few research problems.

8. Bibliography

[1] P. Cichosz, *Systemy uczące się*, WNT, Warszawa, 2007.

[2] S.J. Russel, P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, Englewood Cliffs, New Jersey, 1995.

[3] A. Ameljańczyk, „Analiza wpływu przyjętej koncepcji modelowania systemu wspomagania decyzji medycznych na sposób generowania ścieżek klinicznych", *Raport z realizacji zadania 2. projektu POIG.01.03.01-00-145/08*, WAT, Warszawa, 2009.

[4] B. Kowalczyk, „Przewlekła białaczka szpikowa", *Encyklopedia zdrowia MediWeb.pl*, http://mediweb.pl/diseases/wyswietl_d.php?id=92

[5] *abc Białaczka.pl*, http://abcbialaczka.pl

[6] R. Liu, R. Soetjipto, *Analysis of Three Bayesian Network Inference Algorithms: Variable Elimination, Likelihood Weighting, and Gibbs Sampling*, Berkeley, 2004.

[7] P. Długosz, „Opracowanie koncepcji modułu wspomagania podejmowania decyzji klinicznych w modelu repozytorium z wykorzystaniem metod teorii zbiorów przybliżonych", *Raport z realizacji zadania 3. projektu POIG.01.03.01-00-145/08*, WAT, Warszawa, 2009.

[8] D. Heckerman, *A Tutorial on Learning with Bayesian Networks*, Microsoft Corporation, Redmond, 1995.

[9] R. Niculescu, T. Mitchell, R. Rao, "Bayesian Network Learning with Parameter Constraints", *Journal of Machine Learning Research 7*, 1357−1383, MIT, Boston, 2006.

[10] A. Oniśko i inni, *HEPAR I HEPAR II – komputerowe systemy wspomagania diagnozowania chorób wątroby*, XII Konferencja Biocybernetyki i Inżynierii Biomedycznej, Warszawa, 2001.

[11] S. Andreassen, "MUNIN – An Expert EMG Assistant", *Computer-Aided Electromyography and Expert Systems*, Vol. 21, Elsevier Science Publishers, Amsterdam, 1989.

[12] "Chronic Myelogenous Leukemia", *NCCN Practice Guidelines in Oncology*, http://www.nccn.org/professionals/physician_gls/PDF/cml.pdf

[13] "Acute Myeloid Leukemia", *NCCN Practice Guidelines in Oncology*, http://www.nccn.org/professionals/physician_gls/PDF/aml.pdf

[14] „Ścieżki kliniczne jako dynamiczne środowisko dostępu do informacji medycznej pacjenta", *wersja 0.8 Zintegrowany System Informacji Medycznej o Pacjencie*, Bielsko-Biała – Kraków, 2008.

[15] „Przewlekła białaczka szpikowa", *Wikipedia*, http://pl.wikipedia.org/wiki/Przewlek%C5%82a_bia%C5%82aczka_szpikowa

Koncepcja wykorzystania sieci bayesowskich w module wspomagania decyzji medycznych

M. STRAWA

W artykule przedstawiono koncepcję budowy modułu wspomagania decyzji medycznych, współpracującego z repozytorium ścieżek klinicznych. Składają się na nią: definicja sieci bayesowskich oraz najważniejszych pojęć z nimi związanych, opis wybranego mechanizmu wnioskowania oraz przykład generowania diagnozy w module.

Słowa kluczowe: sieci bayesowskie, sieci przekonań, system wspomagania decyzji medycznych.

Automatyczna budowa semantycznego modelu objawów chorobowych na bazie korpusu słownego

G. SZOSTEK, M. JASZUK, A. WALCZAK
grazyna.szostek@gmail.com

Wydział Cybernetyki Wojskowej Akademii Technicznej w Warszawie
Wyższa Szkoła Informatyki i Zarządzania w Rzeszowie

Opisane w artykule badania dotyczą danych z dziedziny medycyny. Wyniki badań diagnostycznych rejestrowane są na różne sposoby. Mogą mieć postać tabel, wykresów, obrazów. Niezależnie od oryginalnego formatu danych możliwe jest sporządzenie ich opisu słownego, który koncentruje się na opisie zaobserwowanych objawów chorobowych. Opisy takie tworzą korpusy słowne dotyczące poszczególnych technologii diagnostycznych. W podobny sposób zapisywana jest wiedza dotycząca jednostek chorobowych. Ma ona postać korpusów tekstowych, w których zawarte są opisy objawów specyficznych dla poszczególnych schorzeń. Za pomocą narzędzi przetwarzania języka naturalnego możliwe jest automatyczne wydobycie z tekstów modeli semantycznych, opisujących poszczególne technologie diagnostyczne oraz choroby. Pewne utrudnienie stanowi fakt, że wiedza medyczna może zostać zapisana w języku naturalnym na wiele sposobów. Zastosowanie formatu semantycznego pozwala wyeliminować te niejednoznaczności zapisu. W konsekwencji dostajemy ujednolicony model wiedzy medycznej, zarówno od strony wyników technologii diagnostycznych opisujących stan pacjenta, jak i wiedzy dotyczącej jednostek chorobowych. Daje to możliwość dokonania fuzji danych pochodzących z różnych źródeł (danych heterogenicznych) do postaci homogenicznej. Artykuł przedstawia metodę generowania modelu semantycznego wiedzy medycznej, wykorzystującą analizy leksykalne korpusów słownych.

Słowa kluczowe: sieć semantyczna, ontologia, przetwarzanie języka naturalnego.

1. Wprowadzenie

Szybki wzrost ilości informacji i mała skuteczność metod do ich przetwarzania dały początek pracom nad metodami opartymi zarówno na formalnej reprezentacji wiedzy – ontologii, jak i na sieciach semantycznych. W związku z dużą ilością pojęć i zależności między nimi co raz częściej wykorzystuje się automatyczną konstrukcję semantycznego modelu [2], [10]. Podstawowym zasobem wykorzystywanym przez takie metody jest korpus tekstów. Metody przetwarzania języka naturalnego [1], [3], [4], [5] opracowane dla korpusów tekstów w języku angielskim nie mają bezpośredniego przełożenia na języki fleksyjne o swobodnym szyku wyrazów w zdaniu, takie jak język polski. Prace nad konstrukcją sieci semantycznej dla języka polskiego [6], [9] przyczyniły się do powstania automatycznych metod wykrywania leksykalnych relacji semantycznych.

W wielu dziedzinach dane (np. wojskowe, ekonomiczne, medyczne) nie mają jednolitej postaci: tabela, obraz, wykres itd. Każda z nich wymaga stosowania dedykowanych metod przetwarzania i analizy. Od kilku lat sieci semantyczne są wykorzystywane jako narzędzie do jednorodnego zapisu heterogenicznych danych [8]. To podejście odkrywa nowe możliwości przetwarzania danych, zastosowanie tych samych metod analizy do obrazów, tekstów, tabel itd. Format zapisu danych powoduje zmniejszenie ich rozmiaru (np. obrazowych danych), co ma wpływ na szybkość przesyłania danych przez sieć.

Dane medyczne są dobrym przykładem różnorodności zapisu danych. Opisana w pracy budowa modelu semantycznego jest elementem szerszych badań mających na celu dokonanie fuzji danych pochodzących z różnych źródeł (danych heterogenicznych) do postaci homogenicznej. Ten sam objaw może mieć różną postać (fraza w tekście, element tabeli, fragment obrazu, punkt na wykresie), co stanowi problem dla dalszego przetwarzania takiej informacji. Istnieje więc potrzeba zapisu objawów w postaci jednolitej, co daje możliwość zastosowania takich samych metod i narzędzi matematycznych w procesie wspomagania stawiania diagnozy. Podstawą do budowy modelu semantycznego objawów są opisy wyników technologii diagnostycznych i opisy jednostek chorobowych, które tworzą korpus słowny. W artykule zostanie zaprezentowana metoda

Praca została zrealizowana w ramach projektu POIG.02.03.03-00-013/08, dofinansowanego ze środków Europejskiego Funduszu Rozwoju Regionalnego w ramach Programu Operacyjnego „Innowacyjna Gospodarka”.

generowania modelu semantycznego na bazie korpusu słownego. Technika ta zostanie wykorzystana do budowy ontologii poszczególnych technologii diagnostycznych i jednostek chorobowych w celu ujęcia w jednolitą strukturę zapisu wiedzy diagnostycznej.

Układ artykułu jest następujący. W sekcji 2 jest pokazana różnorodność form zapisu danych medycznych i propozycja homogenicznej postaci ich zapisu. Sekcja 3 stanowi krótkie wprowadzenie do modelu formalnego ontologii i sieci semantycznej. Następnie, w sekcji 4 jest szczegółowo omówiony proces wykrywania fraz oznaczających objawy i cech je specyfikujących. Sekcja 5 podsumowuje wyniki uzyskane w sekcji 4. Ogólny opis procesu budowy ontologii jest przedstawiony w sekcji 6.

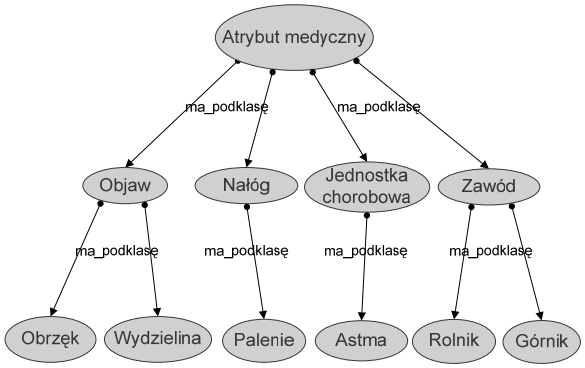
2. Reprezentacja heterogenicznych danych

Opis stanu zdrowia pacjenta i opis jednostki chorobowej składają się z opisów objawów zdiagnozowanych za pomocą różnych technologii. W diagnostyce chorób są m.in. stosowane następujące technologie diagnostyczne (TD):

- badanie podmiotowe
- przedmiotowe
- testy skórne
- gazometria
- RTG
- scyntygrafia
- ultrasonografia
- spirometria
- bronchoskopia
- badanie laboratoryjne itd.

W zależności od TD wyniki mają różną postać. Wyniki mogą być zapisane w postaci tabelarycznej (spirometria, badanie morfologiczne krwi), w postaci wykresu (spirometria), obrazu (RTG), opisu słownego (badanie podmiotowe) itd. Badania obrazowe najczęściej mają dodatkową postać wyniku – opis słowny wykonany przez lekarza specjalistę. Wraz z wykresami mogą być dostarczone najistotniejsze parametry wykresu (w postaci tabeli). Podsumowując, wyniki TD, ze względu na formę zapisu, można podzielić na trzy postacie: tabelaryczną, opisu słownego, pliku grafiki cyfrowej lub analogowej.

W ramach danej TD dokonuje się pomiaru/ obserwacji wielu parametrów związanych ze stanem pacjenta. Można stwierdzić istnienie objawów chorobowych, ale można także zarejestrować dużo dodatkowych faktów dotyczących pacjenta. Przykładami mogą być: przynależność do grupy zawodowej (np. rolnik, górnik), nałogi (np. palenie), istnienie określonych chorób w rodzinie badanego (choroby dziedziczne), wiek itp. Informacje te będą opisywane i przetwarzane przez nasz system w sposób identyczny jak objawy. Zaistniała więc potrzeba zgrupowania wszystkich rodzajów informacji pomocnych przy diagnozowaniu JCH. Zrobimy to przez wprowadzenie osobnej klasy semantycznej, której podklasy będą reprezentowały interesujące nas informacje. Nazwaliśmy tę klasę atrybut medyczny (AM). Rysunek 1 przedstawia przykładową hierarchię klas wywodzących się z klasy Atrybut medyczny.

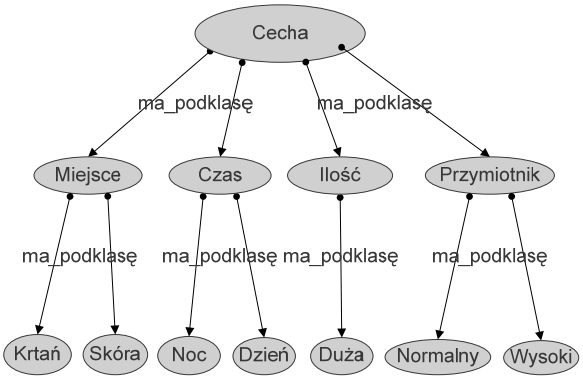


Rys. 1. Przykładowa hierarchia klas wywodzących się z nadrzędnej klasy Atrybut medyczny

Z większością AM można powiązać wiele dodatkowych parametrów, które są cechami charakteryzującymi dany atrybut – cech. Do cech można zaliczyć:

- przymiotnik charakteryzujący atrybut (np. wydzielina: gęsta, zalegająca, obfita itp.)
- miejsce występowania (np. obrzęk: błony śluzowej nosa, powiek, błony śluzowej gardła)
- czas występowania lub nasilania (np. kaszel występujący okresowo lub codziennie, w nocy, między 4–5 rano)
- substancję wywołującą objaw (np. sierść zwierząt, kurz, pyłki roślin, antygeny)
- sytuację, w której występuje objaw (np. stres, zmęczenie, wysiłek fizyczny).

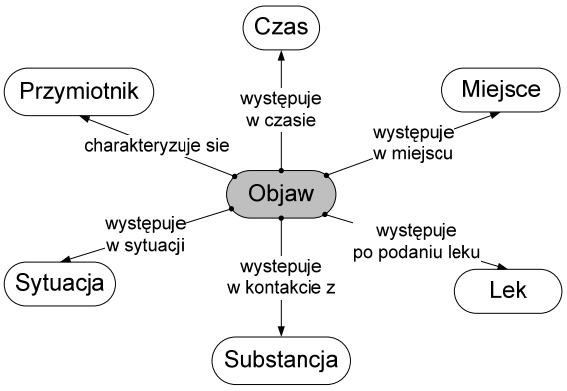
Dla każdej z wymienionych cech można stworzyć osobną klasę semantyczną, która będzie podklasą ogólnej klasy Cecha, rysunek 2.



Rys. 2. Gałąź hierarchii reprezentująca cechy charakteryzujące atrybut medyczny

Struktura hierarchii tworzonej w systemie ma więc dwie główne gałęzie: Atrybuty medyczne i Cechy. Wszystkie pozostałe klasy są podklasami jednej z tych dwóch.

Cechy i AM będą powiązane przez relacje. Ogólną charakterystykę AM reprezentuje fragment ontologii, w którym centralnym węzłem jest konkretny atrybut. Węzeł ten będzie się wiązał z węzłami cech charakteryzujących atrybut. Relacje pomiędzy węzłami będą zależne od typu AM i cechy (rysunek 3). Zarówno AM, jak i ich cechy zostaną zidentyfikowane w tekstach z wykorzystaniem metod przetwarzania języka naturalnego.



Rys. 3. Przykładowe powiązania pomiędzy objawem a cechami go charakteryzującymi

3. Model formalny ontologii i sieci semantycznej

Ontologia

Ontologia jest hierarchicznie i strukturalnie uporządkowanym zbiorem pojęć służących do opisu danej dziedziny:

$$O = \langle C, R, L \rangle . \tag{1}$$

Zbiór C jest zbiorem wszystkich pojęć wykorzystywanych w budowanym modelu.

Element R ontologii O jest zbiorem relacji między pojęciami:

$$R = \{\Re : \Re \subset C \times C\}. \tag{2}$$

Zbiór relacji dzieli się dodatkowo na zbiór relacji strukturalnych oraz relacji hierarchicznych.

Zbiór L jest nazywany leksykonem i jest określony następująco:

$$L = L_C \cup L_R , \tag{3}$$

gdzie:

L_C – zbiór słów stanowiący nazwy dla pojęć, nazywany dalej leksykonem pojęć;

L_R – zbiór słów stanowiący nazwy dla relacji, nazywany dalej leksykonem relacji.

Sieć semantyczna

Z przyjętej definicji ontologii wynika następujące formalne określenie sieci semantycznej:

$$SN^O = \langle I_C^O, I_R^O \rangle , \tag{4}$$

gdzie:

I_C^O – zbiór instancji wszystkich pojęć zdefiniowanych w ontologii O;

I_R^O – zbiór instancji wszystkich relacji pomiędzy pojęciami zdefiniowanymi w ontologii O.

Jeśli $Inst_c$ jest to zbiór instancji pojęcia c, to:

$$I_C^O = \bigcup_{c \in C} Inst_c . \tag{5}$$

Zbiór I_R^O można zapisać jako:

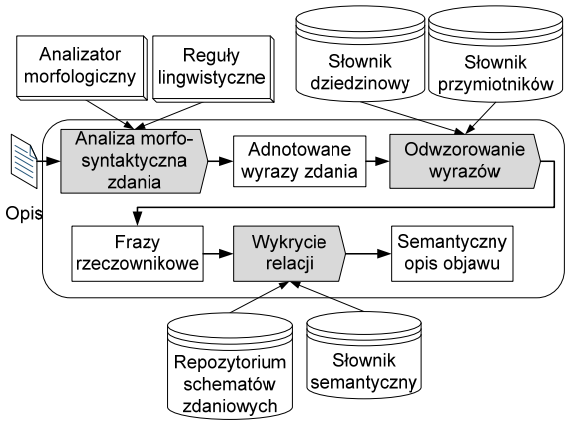
$$I_R^O = \bigcup_{\Re \in R} Inst_{\Re} . \tag{6}$$

Zbiór $Inst_{\Re}$ nazywamy zbiorem instancji relacji $\Re$.

4. Proces tworzenia opisu semantycznego objawów

Opisy wyników diagnostycznych są zapisane w języku naturalnym. Stąd sposób określenia fraz oznaczających objawy i ich cechy jest oparty na metodach przetwarzania języka naturalnego. Na rysunku 4 jest szczegółowo przedstawiony proces budowy opisu semantycznego objawu z pojedynczego zdania. Rezultatem tego procesu jest pojedyncza gałąź sieci

semantycznej, która jest modelem opisującym objaw lub objawy chorobowe. Na podstawie sieci semantycznej będzie budowana ontologia, oznaczymy ją O_{OS}.



Rys. 4. Proces tworzenia opisu semantycznego objawu

Zanim zostanie uruchomiony proces tworzenia opisu semantycznego, konieczne jest stworzenie odpowiednich słowników (słownik dziedzinowy, przymiotników, semantyczny) i zasobów słownikowych (schematy zdaniowe).

Słownik dziedzinowy i słownik przymiotników

Słownik dziedzinowy zawiera rzeczowniki i frazy rzeczownikowe specyficzne dla danego obszaru wiedzy, w omawianym przypadku – medycyny. Słownik jest budowany z wykorzystaniem korpusu tekstów medycznych (opisy wyników TD, opisy JCH, literatura medyczna itd.) i korpusu tekstów o tematyce ogólnej. Słownik przymiotników zawiera przymiotniki charakterystyczne dla rozważanej dziedziny. Przyjmując wcześniej wprowadzone oznaczenia, słownik dziedzinowy i słownik przymiotników stanowią leksykon pojęć L_{C_M} ontologii medycyny O_{M}

Słownik synonimów

Wiele wyrazów zgromadzonych w słowniku dziedzinowym i słowniku przymiotników ma podobne znaczenie, a więc występuje między nimi relacja synonimii. Po zidentyfikowaniu zbiorów takich fraz będą utworzone tzw. synsety, czyli zbiory fraz o tym samym lub zbliżonym znaczeniu, które jednocześnie definiują klasy będące węzłami ontologii. Każdy z synsetów reprezentuje pewne pojęcie. Pojęciem (nazwą synsetu) powinna zostać fraza o największej częstości występowania.

Słownik synonimów jest zbiorem pojęć C_{M} ontologii O_{M}, a słownik dziedzinowy i przymiotników jest leksykonem pojęć L_{C_M}. W hierarchii klas semantycznych przedstawionej na rysunkach 1 i 2 węzły najniższego poziomu zawierają pojęcia ze słownika synonimów.

Słownik semantyczny

Słownik semantyczny grupuje elementy słownika synonimów ($c \in C_M$) poprzez połączenie ich relacją hierarchiczną z odpowiednimi nadklasami. Jak zostało przedstawione w sekcji 2 mamy następujące klasy reprezentujące grupy pojęć: [objaw], [jednostka chorobowa], [czas], [zawód] itd. Te klasy tworzą wyższe poziomy hierarchii klas (rysunek 1 i 2).

Analiza morfosyntaktyczna zdania

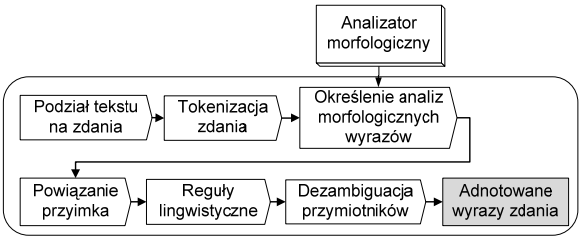
W procesie tworzenia opisu semantycznego brane są pod uwagę opisy wyników tylko z jednej TD lub JCH. Każdy opis składa się ze zbioru zdań. Wykrycie objawu i cech z nim związanych odbywa się w kontekście pojedynczego zdania.

Proces analizy morfosyntaktycznej zdania odbywa się z wykorzystaniem słownika morfologicznego i reguł lingwistycznych. Wynikiem procesu są adnotowane wyrazy zdania – do każdego wyrazu jest przypisany znacznik morfosyntaktyczny.

Schemat procesu analizy składa się z następujących etapów (rysunek 5):

- Tokenizacja

Zdanie jest poddawane tokenizacji, czyli podziałowi na tokeny: słowa, liczby, znaki interpunkcyjne.



Rys. 5. Schemat procesu analizy morfo-syntaktycznej zdania

- Określenie analiz morfologicznych

Dla każdego słowa jest generowany ciąg analiz fleksyjnych przez analizator morfologiczny. Generowane są wszystkie możliwe analizy.

- Powiązanie przyimka

Związek przyimka z rzeczownikiem jest wyrażany za pomocą końcówki rzeczownika charakterystycznej dla przypadku dopuszczalnego w tym połączeniu. Informacja ta umożliwia przeprowadzenie częściowej dezambiguacji niektórych rzeczowników, jak również zaimków, przymiotników i liczebników występujących w parze z przyimkiem.

- Reguła lingwistyczna rzeczownik + rzeczownik (dopełniacz)

W tej części procesu są analizowane wyrazy o klasie gramatycznej rzeczownik. Celem jest ujednoznacznienie kategorii przypadku. Z przeprowadzonych badań wynika, że gdy występują obok siebie w zdaniu „obok siebie" dwa rzeczowniki, np. błona śluzowa oskrzeli, światło oskrzela, świąd skóry itd., to ostatni z nich najczęściej jest w dopełniaczu. Z tej własności skorzystamy przy eliminacji dla tego rzeczownika analiz fleksyjnych zawierających przypadek inny niż dopełniacz.

- Dezambiguacja przymiotników

Zależność między rzeczownikiem i określającym go przymiotnikiem ma wykładnik formalny w postaci końcówek fleksyjnych charakterystycznych dla wspólnego obu wyrazom przypadku i dla wspólnej liczby. Ta własność umożliwia poszukiwanie par – rzeczownik i odpowiadający mu pod względem przypadku i liczby przymiotnik.

W trakcie całego procesu analizy morfosyntaktycznej jest budowana wiedza o podmiocie i orzeczeniu lub podmiotach i orzeczeniach w przypadku zdań złożonych. Opis tego procesu nie mieści się w temacie artykułu.

W trakcie analizy morfosyntaktycznej można wykryć instancje następujących relacji:

charakteryzuje_sie, miejsce_wystąpienia $\in R_{\mathrm{OS}}$,

gdzie R_{OS} – zbiór relacji między pojęciami ontologii O_{OS}.

Instancja relacji występuje między dwoma instancjami pojęć. Nie wszystkie analizowane wyrazy zdania są instancjami pojęć C_{OS} definiowanej ontologii O_{OS}. Część z nich nie występuje w słowniku dziedzinowym, część jest składową fraz rzeczownikowych. Zanim zostaną odrzucone wyrazy nieistotne z punktu widzenia opisu objawów i zanim zostaną wykryte frazy rzeczownikowe, przyjmijmy, że wyrazy zdania są instancjami pojęć ze zbioru C_{x} pewnej ontologii O_{x}.

O wystąpieniu między wyrazami określonej relacji semantycznej możemy wnioskować tylko na podstawie znaczników morfosyntaktycznych przypisanych do wyrazów, ponieważ na tym etapie analizy nie posiadamy informacji semantycznej, jedynie informację morfosyntaktyczną.

Dla relacji *charakteryzuje_się* argumentami są wyrazy z wykrytych fraz rzeczownikowo-przymiotnikowych:

$$\begin{aligned} &Inst_{charakteryzuje\_się} = \{(i,j) \in I_{C_x}^{O_x} \times I_{C_x}^{O_x} : \\ &a(i) \supset \{rzeczownik\}, \\ &a(j) \supset \{przymiotnik\}\}, \end{aligned} \tag{7}$$

gdzie: $a : I_{C_{OS}}^{O_{OS}} \to A$, A – zbiór znaczników morfosyntaktycznych. Relacja między wyrazami pary rzeczownik − rzeczownik w dopełniaczu istnieje, ale na tym etapie nie można określić, jaka to relacja i jaki jest jej kierunek. Dalej instancję relacji $\Re$ będziemy zapisywać:

$$instancja\_relacji(i, j), \tag{8}$$

gdzie: $instancja\_relacji \in Inst_{\Re}$, $i,\ j\ \in I_{C_{OS}}^{O_{OS}}$. Jeśli nie jest znany argument relacji, to zapisujemy „?". Kiedy relacja występuje, ale nie jest znany jej typ, zapisujemy „?relacja". Brak informacji o kierunku relacji, tzn. co jest jej lewym i prawym argumentem, zapisuje się (?lewy_arg, ?prawy_arg). Przynależność do klasy semantycznej zapisujemy jako relację typu hierarchicznego *ma_instancję*:

$$ma\_instancję(klasa\_semantyczna, i), \tag{9}$$

gdzie:

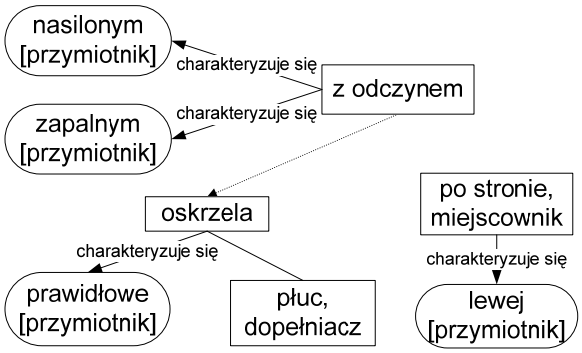
$klasa\_semantyczn \in C_{\mathrm{OS}}, i \in I_{C_{OS}}^{O_{OS}}$.

Przymiotnikom przypisujemy klasę semantyczną [przymiotnik].

Powiązania morfosyntaktyczne między wyrazami wykryte w wyniku analizy można przedstawić za pomocą grafu. W węzłach są umieszczone instancje pojęć, strzałki reprezentują instancje relacji. W węzłach prostokątnych są umieszone rzeczowniki, przymiotniki – w owalnych. Wykryte relacje semantyczne są reprezentowane pełnymi strzałkami. Relacje, które są do wykrycia – strzałkami przerywanymi.

Przykład 1

Dla zdania „Oskrzela płuc są prawidłowe z nasilonym odczynem zapalnym szczególnie po stronie lewej." powstaną poniżej przedstawione grafy.



Rys. 6. Powiązania morfo-syntaktyczne między wyrazami zdania z Przykładu 1

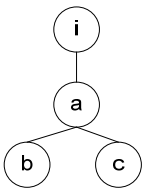
Odwzorowanie wyrazów

Opisany w poprzednim punkcie proces adnotacji opisuje pojedyncze wyrazy. Niektóre wyrazy, gdy występują bezpośrednio obok siebie, tworzą związki mające znaczenie jako całość. Stąd, przechodząc na poziom semantyczny, wymagane jest wykrycie fraz rzeczownikowych. Jak sama nazwa wskazuje, centralnym elementem frazy jest rzeczownik. Proces odwzorowania wyrazów we frazy odbywa się w kilku krokach z wykorzystaniem słownika dziedzinowego. Dla każdego rzeczownika w zdaniu:

$$\forall i \in I_{C_M}^{O_M} : a(i) \supset \{rzeczownik\} \qquad (10)$$

są wykonane kolejne kroki:

1. Utworzenie fraz dla rzeczownika, gdzie fraza zawiera jedno albo więcej słów, na podstawie adnotacji morfologicznej:
   - najpierw dla rzeczownika i są wyszukiwane wyrazy, które są z nim w relacji nie tylko bezpośredniej, ale też pośredniej. Niech zbiór Z_i zawiera powiązane z rzeczownikiem i wyrazy:
   $$Z_i = i \cup \Big\{ j : j \in I_{C_M}^{O_M} \wedge i \neq j \wedge \exists_{r_k \in B \subset I_{R_{OS}}^{O_{OS}}, k=1,\ldots,n, n \leq |B|} \big(r(i,j) \vee r_1(i,\cdot),\ldots,r_k(\cdot,j)\big)\Big\}, \qquad (11)$$
   gdzie: B – zbiór reguł analizowanego zdania.
   - na podstawie zbioru wyrazów Z_i jest generowany zbiór fraz-kandydatów K. Wyraz i występuje w każdej frazie, stąd są tworzone wszystkie podzbiory zbioru $Z_i/\{i\}$ i do tak powstałych fraz jest dodawany wyraz i. Przykładowo, jeśli rzeczownik i występuje w relacji z innymi wyrazami, tak jak pokazano na rysunku 6, to dla niego będą utworzone następujące frazy-kandydaci: *i, ia, ib, ic, iab, iac, ibc, iabc*.



Rys. 7. Przykładowe relacje rzeczownika *i* z innymi wyrazami

   - nie wszystkie elementy zbioru K mogą być frazami. Trzeba odrzucić takie frazy-kandydatów, w których elementy są nieosiągalne z węzła zawierającego rzeczownik i. Dla przykładu z poprzedniego punktu odrzucamy frazy *ib, ic*.
2. Eliminacja ze zbioru K fraz rzeczownikowych, które nie występują w słowniku dziedzinowym. Po wykonaniu tego kroku zbiór K zawiera tylko instancje pojęć ontologii – O_{OS}:
   $$K \subset I_{C_{OS}}^{O_{OS}} \subseteq I_{C_M}^{O_M}. \qquad (12)$$
3. Ze zbioru K jest wybierana najdłuższa dopasowana fraza k.
4. Na podstawie frazy k są modyfikowane odpowiednie reguły w bazie reguł B, tak aby połączyć wyrazy frazy k w jeden węzeł.

Wykrycie relacji

Opis objawu składa się z wielu elementów semantycznie powiązanych ze sobą. Jak zostało wcześniej zaznaczone, takimi elementami są nie tylko charakteryzujące go przymiotniki, ale również elementy opisujące miejsce wystąpienia objawu, czas wystąpienia, czynniki wywołujące objaw itd. W procesie dotychczasowej analizy zostały wykryte relacje morfosyntaktyczne istniejące między wyrazami zdania, tworząc grafy powiązań. Na tym etapie analizy wymagane jest wykrycie relacji semantycznych między grafami reprezentującymi elementy opisu objawu. Proces będzie przebiegał z wykorzystaniem schematów zdaniowych i słow-nika semantycznego.

Schematy zdaniowe

Wydawałoby się, że zdań jest tyle, ile kombinacji wyrażeń, tzn. nieograniczona ilość. Okazuje się, że tak nie jest. Istnienie pewnych zasad i uwarunkowań przy tworzeniu kombinacji wyrazowych powoduje, że można je zapisać w postaci schematów [7]. Kluczowym elementem schematu jest czasownik. Swoimi właściwościami semantyczno-gramatycznymi

decyduje w znacznej mierze o strukturze zdania, w którym występuje. W procesie tworzenia wypowiedzenia wybór wyrazów i ich forma zależą od innych wyrażeń, z którymi wiążą się one gramatycznie i semantycznie. To powoduje, że schematy zdaniowe mogą być bardzo przydatnym narzędziem przy rozwiązaniu takich problemów semantycznych, jak określenie relacji semantycznych, brakujących elementów w zdaniu lub znaczenia wyrazów.

Poszczególne części zdania (podmiot, dopełnienie, okoliczniki itp.) mogą zajmować tylko elementy należące do określonych klas leksykalno-semantycznych. Trzeba także zróżnicować łączliwość obligatoryjną i fakultatywną. Łączliwość obligatoryjna dotyczy składników, które muszą wystąpić przy danym czasowniku. Łączliwość fakultatywna dotyczy składników, które mogą, ale nie muszą, być użyte z danym czasownikiem. Charakterystyka semantyczna składników służy tylko do określenia ograniczeń łączliwościowych.

Algorytm uzgadniania zdania i schematu zdaniowego

Algorytm ma za zadanie znaleźć schemat zdaniowy dla danego zdania, wygenerować oczekiwania w odniesieniu do brakujących składowych i określić znaczenie wyrazów wieloznacznych (w sensie semantycznym) w kontekście danego zdania.

Każdy wyraz przechodzi przez analizę morfosyntaktyczną. Szczególną uwagę zwraca się na czasowniki i rzeczowniki (frazy rzeczownikowe). Czasownik pozwoli dotrzeć do właściwych schematów, a rzeczowniki pomogą wybrać jeden z nich.

Etap pierwszy polega na przypisaniu instancji do klas semantycznych. W tym celu zostanie wykorzystany słownik semantyczny. Dla rzeczowników z pary rzeczownik-rzeczownik w dopełniaczu sprawdza się: jeśli rzeczownik w dopełniaczu ma klasę [element_anatomii], to wiążemy oba rzeczowniki relacją *miejsce_wystąpienia* i rzeczownik w dopełniaczu zapisujemy jako drugi argument relacji.

Po określeniu w zdaniu czasownika, wybierane są ze zbioru schematów zdaniowych te schematy, w których dany czasownik występuje. Dokonuje się wstępnego wyboru schematów z listy, wykorzystując przy tym frazy rzeczownikowe.

Drugi etap polega na rozwiązywaniu niejednoznaczności. Jeśli pierwszy etap zakończył się na wybraniu z listy jednego schematu i wszystkie jego składowe występują w zdaniu, to algorytm kończy działanie. Brak w zdaniu pewnych fraz rzeczownikowych wymaganych przez schemat nie oznacza, że wybraliśmy niewłaściwy schemat. Taka sytuacja może wystąpić, kiedy rzeczownik został już wspomniany w poprzednich zdaniach lub będzie o nim mowa w kolejnych zdaniach. W celu ustalenia, jaki to jest rzeczownik, w pierwszym przypadku trzeba przeprowadzić analizę semantyczną dotychczasowych wyników, w drugim – wygenerować oczekiwanie.

Zbiór czasowników, które mogą pojawić się w opisach JCH i TD jest ograniczony. Dla większości z nich liczba schematów zdaniowych ogranicza się do jednego lub dwóch.

Niektóre czasowniki mogą pojawić się w wielu schematach zdaniowych, np. czasownik *być*. Samo określenie przypadku dla rzeczowników może być niewystarczającym kryterium wyboru właściwego schematu. W takim przypadku pomocny będzie słownik semantyczny.

Połączenie informacji morfo-syntaktycznej z semantyczną umożliwi ujednoznacznienie wyboru schematu zdaniowego, a to spowoduje rozwiązanie problemu wieloznaczności semantycznej.

Przykład 2

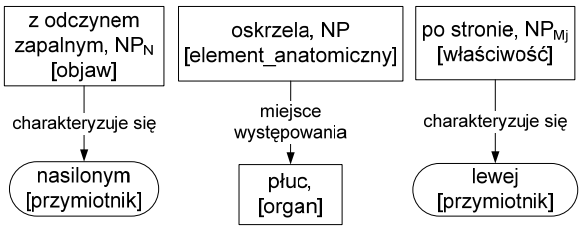
W zdaniu – *Oskrzela płuc są prawidłowe z nasilonym odczynem zapalnym szczególnie po stronie lewej* – występuje czasownik *jest* jako orzeczenie. Poniżej są przedstawione wybrane schematy zdaniowe dla czasownika *jest*:

1. **NP – NP_N+(NP_{Mj})**
   NP→ [element_anatomii]
   NP_N→ [objaw]
   NP_{Mj}→ [właściwość][element_anatomii]
   R: miejsce_wystąpienia(NP_N, NP)
   charakteryzuje_się(NP, NP_{Mj})
2. **NP – NP_N + NP_D**
   NP→ [element_anatomii]
   NP_N → „lokalizacja”
   NP_D→ [objaw]
   R: miejsce wystąpienia(NP_D, NP)
3. **NP – Adj**
   NP→ [objaw]
   NA → [właściwość]
   R: charakteryzuje_się(NP, NA)
4. **NP – NP_N**
   NP → [objaw]
   NP_N → „objaw”

Objaśnienia do schematów:
NP: fraza rzeczownikowa;

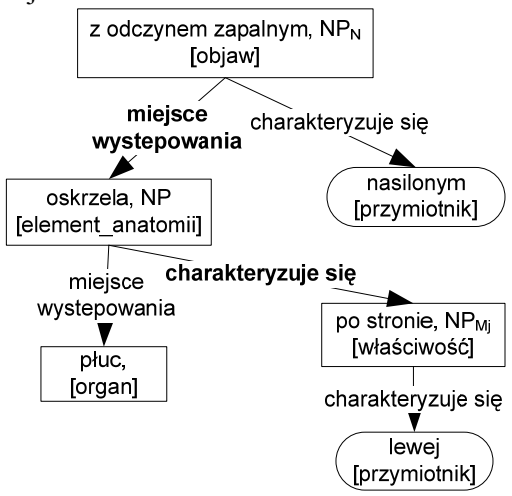
$NP_{D,B,C,N,Mj}$: litery u dołu fraz rzeczownikowych oznaczają ich przypadek (dopełniacz, biernik, celownik, narzędnik, miejscownik);
Adj: przymiotnik;
— pozycja czasownika w schemacie zdaniowym;
+ między składnikami oznacza ich łączenie bez implikacji szyku w aktualnym zdaniu;
( ) fakultatywność składników lub grupy składników (tj. możliwość ich pominięcia);
→ strzałka odsyła do charakterystyki semantycznej;
[ ] dla fraz rzeczownikowych są podane klasy semantyczne;
R: typ relacji.

W korzeniach już wykrytych relacji (w trakcie analizy morfosyntaktycznej) mamy frazy rzeczownikowe: *z odczynem zapalnym, oskrzela, po stronie*.



Rys. 8. Relacje wykryte w trakcie analizy morfo-syntaktycznej

Fraza w narzędniku występuje w schematach 1 i 4, fraza w miejscowniku (może wystąpić, ale nie musi) – tylko w schemacie 1. Podmiot zdania o klasie semantycznej [element_anatomii] pojawia się w schematach 1 i 3. Po sprawdzeniu dopasowania klas semantycznych, wybrany zostaje schemat 1. Uzupełniamy grafy o nowe relacje:



Rys. 9. Graf reprezentujący opis semantyczny zdania z przykładu 2

5. Sieć semantyczna

Produktem procesu wykrywania fraz oznaczających objawy i cech specyfikujących je jest sieć semantyczna stworzona na podstawie tekstu. Każde ze zdań w tekście dostarcza nam pojedynczej gałęzi do struktury sieci. Na podstawie sieci semantycznej będzie zbudowana ontologia – O_{OS}. Z wprowadzonego wcześniej formalnego określenia sieci semantycznej wiadomo, że sieć definiują dwa zbiory. Dla budowanej ontologii O_{OS} będzie to:

$$SN^{O_{OS}} = \left\langle I_{C_{OS}}^{O_{OS}}, I_{R_{OS}}^{O_{OS}} \right\rangle, \qquad (13)$$

gdzie: $I_{C_{OS}}^{O_{OS}}$ – zbiór instancji pojęć ontologii O_{OS} budowanej dla TD lub JCH; zbiór składa się z wyrazów (fraz) występujących we wszystkich opisach analizowanych w ramach danej TD lub JCH; trzeba zaznaczyć, że $I_{C_{OS}}^{O_{OS}} \subseteq I_{C_M}^{O_M}$;

$I_{R_{OS}}^{O_{OS}}$ – zbiór instancji relacji pomiędzy pojęciami C_{OS} ontologii O_{OS}; zbiór zawiera wszystkie relacje wykryte w trakcie analizy zdań TD lub JCH.

6. Tworzenie ontologii

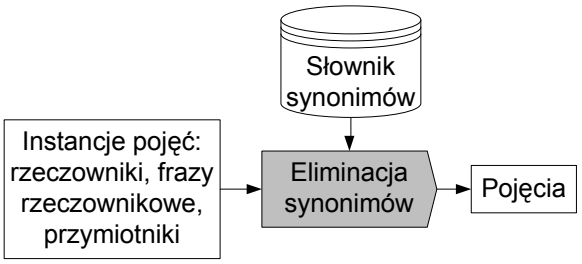
Tworzenie ontologii odbywa się poprzez eliminację synonimów w sieci semantycznej i przekształcenie sieci w ontologię.

Treść o tym samym znaczeniu można zapisać na wiele sposobów, używając przy tym wyrazów lub fraz o zbliżonym znaczeniu. Etap eliminacji synonimów polega na zastąpieniu takich fraz na pojęcia ze słownika synonimów.

$$\forall i \in I \subset I_{C_{OS}}^{O_{OS}}, V_{C_{OS}}^{O_{OS}}(i) \in C_{OS}, \qquad (14)$$

gdzie: $V_{C_{OS}}^{O_{OS}} : I_{C_{OS}}^{O_{OS}} \to C_{OS}$;

$I_{C_{OS}}^{O_{OS}}$ – zbiór instancji wszystkich pojęć C_{OS} zdefiniowanych w ontologii O_{OS};
I – zbiór instancji pojęć w analizowanym zdaniu.



Rys. 10. Schemat eliminacji synonimów

Sieć semantyczną zredukowaną do sieci pojęć przekształca się w ontologię poprzez eliminację wielokrotnych powtórzeń tych samych związków semantycznych.

7. Podsumowanie

Automatyczna budowa semantycznego modelu objawów chorobowych na bazie korpusu słownego jest narzędziem do zbudowania ontologii technologii diagnostycznej lub jednostki chorobowej. Celem konstrukcji ontologii TD jest akwizycja danych diagnostycznych o pacjencie. Ontologia posłuży do budowy interfejsu, poprzez który możliwe będzie wprowadzanie i semantyzacja tych danych. Ontologia JCH jest budowana w celu ujednolicenia opisu jednostek chorobowych. Dzięki zastosowaniu jednolitego formatu do opisu JCH i stanu pacjenta będzie możliwe przeprowadzenie procesu diagnozowania z wykorzystaniem wszelkich możliwych danych. Dalsze prace będą związane z poprawieniem wyników analizy morfosyntaktyczniej, rozbudową systemu o możliwość automatycznej identyfikacji schematów zdaniowych. Do rozważenia jest opracowanie metody wstępnej selekcji zdań ze względu na interesujące nas informacje. Umożliwiłoby to odrzucenie zdań, które nie zawierają informacji istotnej z punktu widzenia opisu objawów.

8. Bibliografia

[1] C. Burgess, „Representing and resolving semantic ambiguity: A contribution from high-dimensional memory modeling”, in: Gorfein, D.S. (Ed.), *On the Consequences of Meaning Selection: Perspectives on Resolving Lexical Ambiguity*, APA Press, 2001.

[2] H. Chen, K.J. Lynch „Automatic construction of networks of concepts characterizing document database”, *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 22, No. 5, 885–902 (1992).

[3] Z.S. Harris, „Mathematical Structures of Language”, *Interscience Publishers*, New York, 1968.

[4] M.A. Hearst, „Automatic Acquisition of Hyponyms from Large Text Corpora”, *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, 1992.

[5] K. Lund, C. Burgess, „Producing high-dimensional semantic spaces from lexical co-occurence”, *Behavior Research Methods, Instrumentation and Computers*, 28, 203–208 (1996).

[6] M. Piasecki, M. Derwojedowa, P. Koczan, A. Przepiórkowski, S. Szpakowicz, M. Zawisławska, „Półautomatyczna konstrukcja Słowosieci” URL www.plwordnet.pl/main. Strona domowa projektu (2007).

[7] K. Polański (red.) *Słownik syntaktyczno-generatywny czasowników polskich,* t. 1–7, Kraków, 1980–1993.

[8] J. Rohmer, ”The Case for Using Semantic Nets as a Convergence Format for Symbolic Information Fusion in NATO”, *RTO-MP-IST-040 Information Systems Technology Panel (IST) Symposium on Military Data and Information Fusion*, Prague, Czech Republic, 2003.

[9] Słowosiec. Witryna WWW projektu. URL http://www.plwordnet.pwr.wroc.pl/main, (2007).

[10] P. Velardi, P. Fabriani, M. Missikoff, „Using text processing techniques to automatically enrich a domain ontology”, in: *Proceedings of the international Conference on Formal Ontology in Information Systems*, FOIS '01, ACM, 270–284, New York, 2001.

Automatic construction of a semantic model of disease symptoms based on text corpus

G. SZOSTEK, M. JASZUK, A. WALCZAK

The research described in article refers the medical data. Descriptions of diagnostic technologies results and descriptions of diseases form the text corpus. The corpus is the basis for building a semantic model of symptoms. A specific symptom can be written in the natural language in many ways, which is a problem for further processing of such information. There is a need to record symptoms in a uniform format. Such format allows for application of the same methods and mathematical tools to support the process of diagnosis. The paper presents method of generating a semantic model based on text corpus. Construction of the model is a part of the research, which aims to make the fusion of data from different sources (heterogeneous data) into homogeneous form.

Keywords: semantic network, ontology, natural language processing.