Marcin Mazurek[1]

Indeksowanie zawartości repozytorium danych medycznych pojęciami ontologii

1. Wstęp

Obecność technologii informatycznej w medycynie skutkuje gromadzeniem danych, które mogą zostać wykorzystane do wspomagania decyzji klinicznych. Aby jednak wykorzystać te dane, konieczna jest ich transformacja do modelu akceptowalnego przez narzędzia analityczne oraz warstwa semantyczna, umożliwiająca ich poprawną interpretację znaczeniową. Rolę taką pełniły hurtownie danych, wymuszające unifikację formatów i modelu w momencie zapisu do bazy danych. Wobec wyzwań związanych z przetwarzaniem danych masowych: szybkości napływu, wolumenu danych i braku struktury tradycyjne podejście prowadzi do wstrzymania części danych, niespełniających kryteriów poprawności, do chwili zaimplementowania odpowiednich mechanizmów zapewnienia ich jakości.

Alternatywnym sposobem przechowywania danych może być baza NoSQL, która umożliwia zapisanie wszystkich napływających danych dzięki elastycznemu modelowi, pozwalającemu zapisać dane źródłowe bez zmiany schematu. Jednocześnie baza ta umożliwia późniejsze dołączenie ustandaryzowanych definicji danych oraz konwersję do wymaganego modelu danych.

Bazując na koncepcji repozytorium danych medycznych wykorzystującego cechy baz NoSQL[2], w artykule przedstawiono prototypową implementację bazy danych indeksowanej pojęciami ontologii medycznej SNOMED CT, zbudowaną z wykorzystaniem MongoDB. Do bazy zaimplementowane zostały procesy ładowania przykładowych wyników badań oraz interfejs pobierania zawartości z wykorzystaniem pojęć zdefiniowanych w ontologii.

---

[1] Wojskowa Akademia Techniczna w Warszawie, Wydział Cybernetyki.

[2] M. Mazurek, *Architektura systemu wspomagania decyzji medycznych wykorzystująca technologię przetwarzania danych Big Data*, „Roczniki Kolegium Analiz Ekonomicznych" 2014, z. 35, s. 257–271.

Przedstawione rozwiązanie wpisuje się w kierunki badań opisywane w literaturze, zmierzające do integracji heterogenicznych źródeł danych w postaci jednolitego interfejsu wystawionego użytkownikowi końcowemu, wykonującemu zaawansowane analizy eksploracji danych[3]. W literaturze można odnaleźć propozycje wykorzystania zarówno architektury Big Data[4], jak też samej ontologii[5].

2. Ontologia SNOMED CT

Ontologia może służyć jako platforma do formalnej budowy informacji, preferencji i wiedzy. Wykorzystuje się w niej kategoryzację, czyli przyporządkowanie symbolu/terminu do określonej grupy obiektów i hierarchizację: umiejscowienie określonej klasy w hierarchicznej strukturze klas opisujących daną dziedzinę.

SNOMED CT to systematycznie zorganizowany, przetwarzany komputerowo zbiór pojęć medycznych dostarczający kody, pojęcia, synonimy i definicje używane w medycynie dla celów klinicznych. Może być wykorzystany do opisania historii przypadku medycznego, rozprzestrzeniania się epidemii czy szczegółów zabiegu ortopedycznego. Jest rozwijany i licencjonowany przez IHTSDO (*International Health Terminology Standards Development Organisation*) – Międzynarodową Organizację ds. Rozwoju Standardów Terminologii Medycznej.

SNOMED CT składa się z ponad 300 tys. pojęć oraz 1,4 mln opisów i jest kluczowym elementem przy tworzeniu elektronicznej dokumentacji medycznej (EDM, ang. EHR – *Electronic Health Record*). Oferuje także mapowania (mapy krzyżowe) dla innych nomenklatur medycznych, jak ICD-10[6] czy LOINC[7].

---

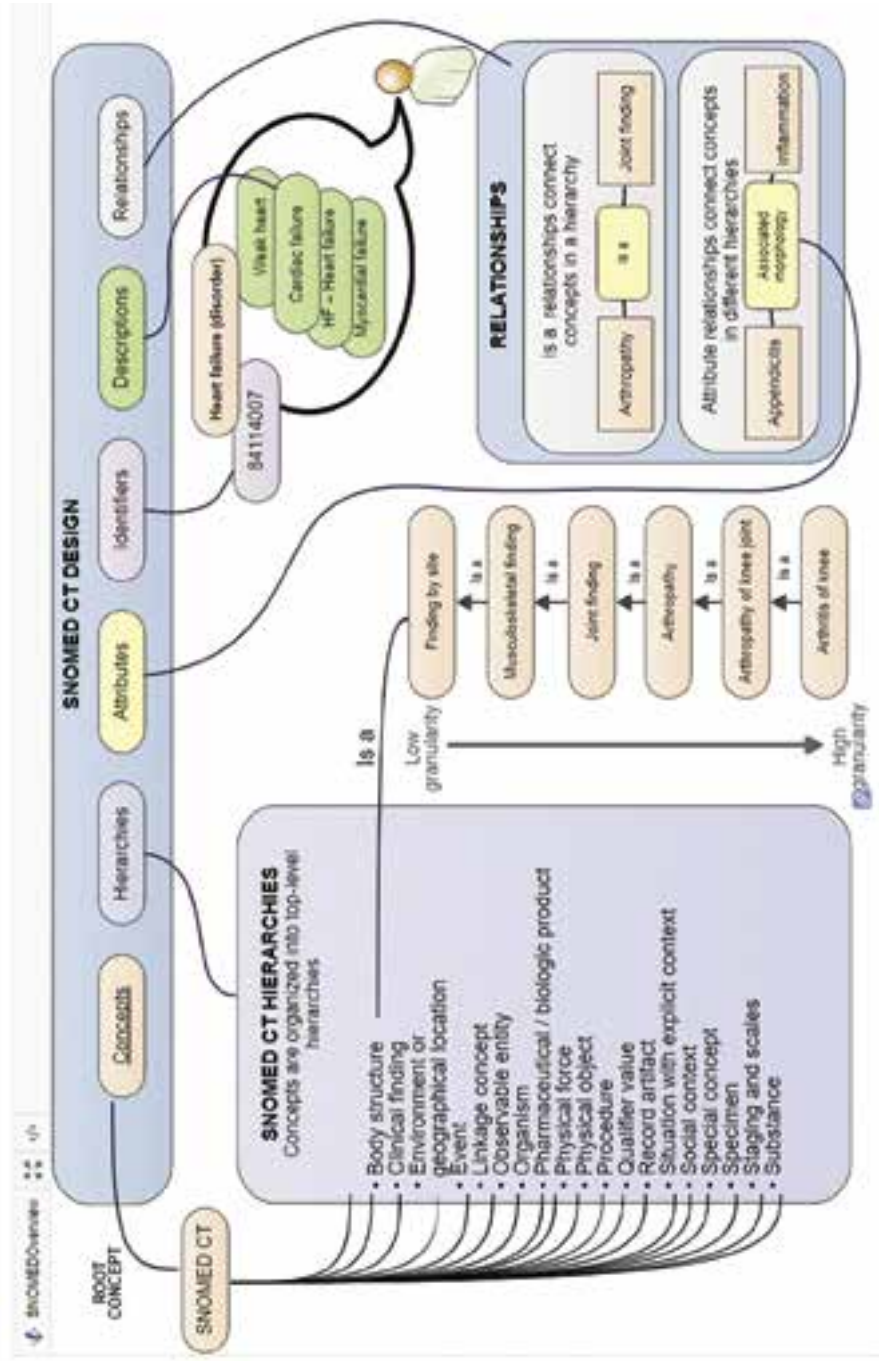
[3] M.H. Tekieh, B. Raahemi, *Importance of Data Mining in Healthcare: A Survey*, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Paris, August 25–28, 2015.

[4] N. Cassavia, M. Ciampi, G. De Pietro, E. Masciari, *A Big Data Approach For Querying Data in EHR Systems*, IDEAS '16: Proceedings of the 20th International Database Engineering & Applications Symposium, Montreal, July 11–13, 2016; K. Batko, *Możliwości wykorzystania technologii Big Data w ochronie zdrowia*, „Roczniki Kolegium Analiz Ekonomicznych" 2016, z. 42, s. 267–282.

[5] R. Khare, Y. An, J. Li, I.L. Song, X.Hu, *Exploiting semantic structure for mapping user-specified form terms to SNOMED CT concepts*, Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (IHI '12), Miami, January 28–30, 2012, ACM, New York 2012, s. 285–294.

[6] http://apps.who.int/classifications/icd10/browse/2016/en (data odczytu: 20.11.2017).

[7] https://loinc.org/ (data odczytu: 20.11.2017).



Rysunek nr 1. Budowa ontologii SNOMED CT

Źródło: SNOMED CT Starter Guide

Podstawowymi elementami ontologii SNOMED CT, przedstawionymi schematycznie na rysunku nr 1, są:

- Pojęcia (ang. *concepts*) – reprezentują różnego rodzaju byty medyczne, pogrupowane w hierarchiczne kategorie, odpowiadające: procedurom medycznym, anatomii, diagnozom itd. Każde pojęcie ma unikalny identyfikator numeryczny oraz opis (ang. *descriptions*) zrozumiały dla człowieka. Każdy koncept może mieć kilka alternatywnych opisów, w zależności od istniejących synonimów bądź wersji językowych.
- Relacje (ang. *relationships*) – łączą pojęcia, wskazując jednocześnie na rodzaj tego związku, np. uogólnienie/uszczegółowienie. Przykładami relacji są również miejsce występowania (ang. *finding site*) lub związek przyczynowo-skutkowy (ang. *causative agent*). W ontologii SNOMED typ relacji, w jakiej może pozostawać pojęcie, jest jego atrybutem. Lista typów relacji (a więc atrybutów) pojęć jest charakterystyczna dla kategorii pojęcia.
- Zestawy Referencyjne (ang. *reference sets*) – grupują kody pojęciowe i opisy w mapy krzyżowe, będące referencjami do innych terminologii i standardów, na przykład wspomnianej bazy pojęć LOINC.

Zawartość ontologii SNOMED jest udostępniona poprzez aplikację WWW umożliwiającą wyszukiwanie pojęć i ich powiązań, lub poprzez interfejs programistyczny API. Możliwe jest również pobranie zawartości ontologii, przy zachowaniu warunków licencyjnych. Takie formy udostępnienia bazy pojęć umożliwiają wykorzystanie jej w systemach przetwarzania danych medycznych.

3. Dokumentowa baza danych MongoDB

MongoDB jest bazą NoSQL przeznaczoną do przechowywania danych zgodnie z modelem dokumentów JSON[8]. Ma możliwość skalowania dla przetwarzania danych masowych w oparciu o platformę Hadoop[9], posiadając jednocześnie zalety relacyjnych baz danych: indeksy oraz rozbudowany język definiowania zapytań.

Z punktu widzenia zastosowań medycznych, cechą wyróżniającą MongoDB jest elastyczny model danych. Podstawową jednostką przechowywania danych w MongoDB jest dokument, na który składają się dwójki uporządkowanych pól

[8] https://www.json.org/ (data odczytu: 20.11.2017).
[9] https://www.mongodb.com/hadoop-and-mongodb (data odczytu: 20.11.2017).

(ang. *field*) i ich wartości (ang. *value*). Poprzez dopuszczenie zagnieżdżonych dokumentów oraz tablic podejście dokumentowe pozwala na reprezentację złożonych hierarchicznych związków za pomocą pojedynczego rekordu. Ten sposób reprezentacji danych jest bardziej zbliżony do modelu realizowanego przez większość języków zorientowanych obiektowo niż model relacyjny.

Kolekcja jest kontenerem przechowującym grupę dokumentów. Jeśli dokument jest w MongoDB odpowiednikiem wiersza w relacyjnej bazie danych, to kolekcja może być rozpatrywana jak odpowiednik relacyjnej tabeli. Kolekcje natomiast nie mają narzuconego schematu danych, dokumenty przechowywane w obrębie tej samej kolekcji mogą nie tylko różnić się typami wartości przechowywanych danych, ale także posiadać całkowicie inny zestaw kluczy.

Szczególnym typem kolekcji jest kolekcja o ograniczonym rozmiarze (ang. *capped collection*). Kolekcje o ograniczonym rozmiarze zachowują chronologię dodawanych dokumentów (również podczas uzyskiwania wyników poprzez zapytania), a jeśli podczas próby dodania kolejnego dokumentu którekolwiek z ograniczeń rozmiaru kolekcji zostałoby przekroczone, usunięty zostaje najstarszy dokument, tworząc tym samym miejsce dla nowego dokumentu. Kolekcje te stanowią cykliczny bufor, z którego dane są automatycznie pobierane przez dedykowane kursory (ang. *tailable cursor*). Kursor pobierający dane z kolekcji o ograniczonym rozmiarze, w odróżnieniu od kursorów domyślnych, nie jest zamykany po wyczerpaniu wyników, lecz nieustannie obserwuje kolekcję w oczekiwaniu na dodanie nowych danych do kolekcji. Kiedy kursor wyczerpie wyniki, MongoDB blokuje wątek zapytania do czasu pojawienia się nowych danych. Gdy to nastąpi, wątek jest powiadamiany, zostaje odblokowany i zwraca kolejną porcję danych.

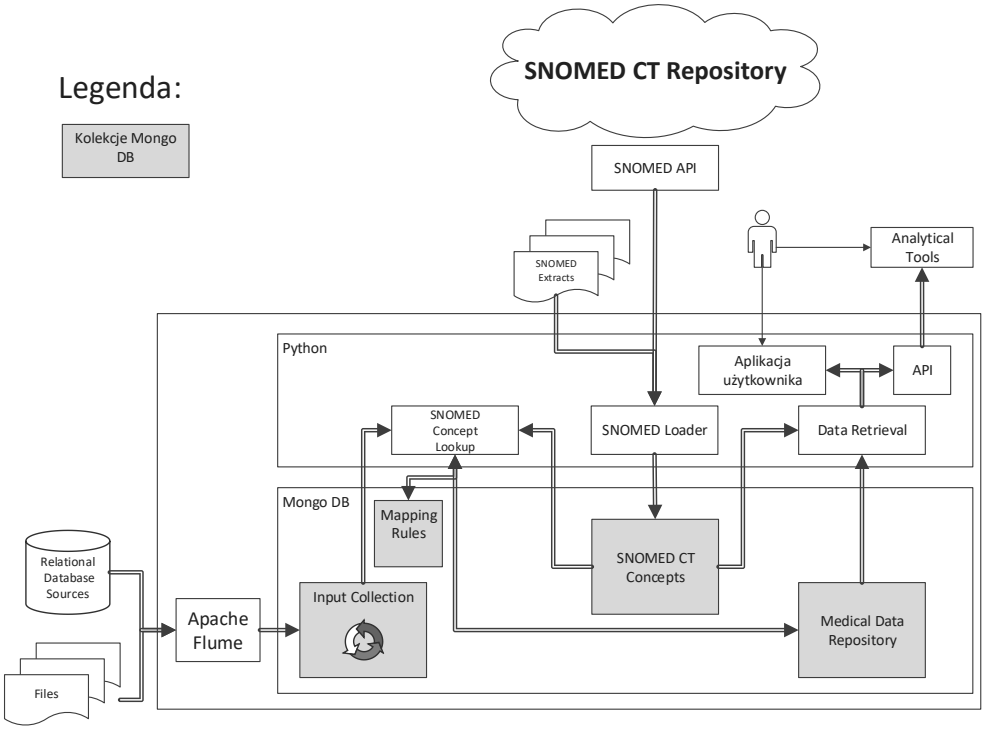
4. Architektura rozwiązania

Wykorzystując wymienione powyżej cechy systemu zarządzania nierelacyjną bazą danych MongoDB, zbudowane zostało prototypowe repozytorium danych medycznych. Jego architekturę przedstawia rysunek nr 2.

4.1. Proces gromadzenia danych

Automatyzację procesu ładowania danych osiągnięto dzięki zastosowaniu mechanizmów strumieniowego ładowania danych udostępnionych przez Apache

Flume[10]. Możliwe jest podłączenie jako źródeł relacyjnych baz danych, jak również plików tekstowych oraz usług sieciowych Web Service. Na tym etapie przetwarzania priorytetem jest duża przepustowość procesu przenoszenia danych, natomiast weryfikacja znaczeniowa odbywa się w postaci procesu wsadowego w późniejszych etapach.



Rysunek nr 2. Architektura repozytorium danych medycznych

Źródło: opracowanie własne

4.2. Składnice repozytorium

W skład repozytorium wchodzą następujące kolekcje dokumentów:

- Input Collection – kolekcja o ograniczonym rozmiarze, mająca charakter cyklicznego bufora. Są do niej ładowane źródłowe dane, przetransformowane w dokumenty JSON. Proces łączenia z pojęciami ontologii SNOMED CT

[10] https://flume.apache.org/ (data odczytu: 20.11.2017).

uruchamiany jest automatycznie dla każdego nowo dodanego dokumentu za pomocą kursora obserwującego stan kolekcji.

- Medical Data Repository – główna kolekcja systemu, w której przechowywane są dokumenty JSON będące wynikiem procesu mapowania (indeksowania). Dane z tej kolekcji udostępniane są użytkownikom. Do tego repozytorium trafiają wszystkie dane – niezależnie od wyniku działania procesu łączenia danych.
- SNOMED Collection – w tej kolekcji przechowywana jest ontologia SNOMED CT, wykorzystywana podczas procesu indeksowania. Każdy koncept ontologii jest przechowywany jako jeden dokument, w którego skład wchodzi kod pojęciowy (ang. *concept code*) oraz opis (ang. *description*). Jako że zapytania wybierające dane z powyższej kolekcji są wykonywane wielokrotnie, podczas procesu mapowania oraz zadawania zapytań do repozytorium utworzony został indeks tekstowy. Budowa indeksu daje również możliwość wyszukiwania konceptów ontologii za pomocą ciągów tekstowych, dzięki czemu użytkownik jest zwolniony ze znajomości kodów pojęciowych podczas korzystania z interfejsu do zadawania zapytań.
- Mapping Rules – kolekcja przechowująca reguły łączenia danych. Jest to pomocnicza baza danych algorytmów łączenia danych źródłowych z definicjami pojęć. Jej zawartość może być edytowana przez użytkownika lub automatycznie uzupełniana jako wynik działania algorytmów łączenia danych.

4.3. Indeksowanie dokumentów

Zadaniem procesu indeksowania, czyli łączenia, jest określenie zawartości dokumentu JSON w sposób jednoznaczny z wykorzystaniem terminologii SNOMED CT, służącej do opisywania danych pacjenta dla celów klinicznych. Dane poddane temu procesowi mogą być dalej wykorzystywane do analiz predykcyjnych, asocjacji, sekwencji czy wyszukiwania skupień.

Proces mapowania jest wykonywany dla każdego nowo dodanego do kolekcji wejściowej dokumentu. Każda para klucz–wartość wchodząca w skład dokumentu jest przetwarzana oddzielnie. Ciąg znaków reprezentujący klucz jest czynnikiem pozwalającym na interpretację pola wartości, dlatego pełni on podstawową rolę podczas mapowania. Mapowanie bazuje na wyszukiwaniu w opisach tekstowych konceptów ontologii wspomnianego ciągu znaków będącego kluczem. Identyfikatory konceptów odpowiadające wartości klucza są dodawane do przetwarzanego dokumentu w formie tablicy. Przetworzone dokumenty będące wynikiem procesu mapowania przenoszone są do kolekcji Medical Data Repository.

Moduł indeksowania dokumentów wykorzystuje do działania reguły, zapisane w bazie reguł Mapping Rules. Przykładem takiej reguły może być: jeżeli klucz jest równy „WBC" wtedy wartość odnosi się do pojęcia 767002, czyli „White blood cell count". Reguły mogą być dodawane ręcznie przez użytkownika bądź na kolejnym etapie rozwoju systemu uzupełniane w procesie uczenia się systemu, wykorzystującego metody uczenia maszynowego. Algorytm łączenia może odwoływać się również do pól wartości i na ich podstawie „dogadywać" ich znaczenie.

4.4. Interfejs do zadawania zapytań

Dane zgromadzone w Medical Data Repository są udostępniane użytkownikowi poprzez moduł zapytań Data Retrieval, który umożliwia zadawanie pytań o dane z wykorzystaniem pojęć ontologii. Moduł ten udostępnia usługi pobierania danych w formacie dwuwymiarowej tabeli danych lub szeregów czasowych. Innym sposobem dostępu jest wykorzystanie bezpośrednio interfejsu programistycznego bazy danych Mongo DB.

4.5. Graficzny interfejs użytkownika

Aplikacja użytkownika jest pomocniczym interfejsem umożliwiającym zadawanie zapytań do Medical Data Repository. W systemie wykorzystywanym produkcyjnie jej znaczenie będzie ograniczone do testowania poprawności działania algorytmów łączenia danych i rzadkich przypadków sięgania do pojedynczych dokumentów źródłowych.

Document table

Patient_ID	date	139576007	38082009	125605004	390396009
0	2017-01-21...		0.0		30.0
1	2017-01-21...		13.0		35.0
2	2017-01-21...		19.0		13.0
3	2017-01-21...	True			
4	2017-01-21...			True	
6	2017-01-21...		15.0	True	

Rysunek nr 3. Przykładowy rezultat zwrócony w aplikacji użytkownika

Źródło: opracowanie własne

5. Podsumowanie

Przedstawione rozwiązanie integruje ontologię dla danych medycznych SNOMED CT z danymi przechowywanymi w repozytorium NoSQL. Integracja polega na zbudowaniu odwzorowania obiektów przechowywanych w repozytorium NoSQL na pojęcia zdefiniowane w ontologii. Dzięki temu możliwe jest zadawanie zapytań i pobieranie danych za pomocą uniwersalnych pojęć znanych lekarzom i naukowcom, bez konieczności znajomości struktur danych. Wynik zapytania, które może łączyć dane z wielu systemów źródłowych o różnych schematach danych, generowany jest w formie tabelarycznej, gdyż taka forma danych wymagana jest przez większość narzędzi eksploracji danych.

Przedstawione rozwiązanie ma charakter prototypu, potwierdzającego poprawność architektury technicznej i możliwość integracji komponentów w sposób opisany w artykule. Można wskazać dwa obszary badawcze rozwoju funkcjonalności komponentów, które w znaczący sposób zwiększą przydatność przedstawionego systemu:

- wnioskowanie o znaczeniu pola nie tylko w oparciu o etykietę, ale również przesyłane wartości (heurystyczne algorytmy łączenia danych oparte na metodach sztucznej inteligencji),
- zakres funkcjonalności interfejsów udostępniających dane i ich integracja z API repozytorium ontologii, tak aby można było wykorzystywać ją nie tylko w roli słownika pojęć, ale również jako źródło informacji o powiązaniach pomiędzy danymi.

Przedstawiona koncepcja zakłada integrację danych w warstwie pojęciowej oraz oddzielenie procesów pozyskiwania i gromadzenia danych od procesów integracji semantycznej. Takie podejście maksymalizuje korzyści analityków w obszarze szybkości dostępu do danych – dane, których znaczenie udało się opisać przez koncepty ontologii są automatycznie dostępne dla narzędzi eksploracji danych.

Bibliografia

An Y., Mylopoulos J., Borgida A., *Building semantic mappings from databases to ontologies*, Proceedings of the 21st National Conference on Artificial Intelligence (AAAI '06), Boston, July 16–20, 2006, t. 2, AAAI Press, 2016.

Batko K., *Możliwości wykorzystania technologii Big Data w ochronie zdrowia*, „Roczniki Kolegium Analiz Ekonomicznych" 2016, z. 42, s. 267–282.

Cassavia N., Ciampi M., De Pietro G., Masciari E, *A Big Data Approach For Querying Data in EHR Systems*, IDEAS '16: Proceedings of the 20th International Database Engineering & Applications Symposium, Montreal, July 11–13, 2016, ACM 2016.

Chodorow K., Dirlof M., *MongoDB the definitive guide*, O'Reilly, Gravenstein Highway North Sebastopol, 2010.

Khan A., Cohen R., Fu L., Doucette J., Jin C., *An ontological approach to data mining for emergency medicine*, Proceedings of the 40th Annual Meeting Northeast Decision Sciences Institute Conference, Northeast Decision Sciences Institute, Montreal, April 14–16, 2011.

Khare R., An Y., Li J., Song I.Y., Hu X., *Exploiting semantic structure for mapping user-specified form terms to SNOMED CT concepts*, Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (IHI '12), Miami, January 28–30, 2012, ACM, New York 2012.

Li X., Morie P., Roth D., *Semantic Integration in Text: From Ambiguous Names to Identifiable Entities*, „AI Magazine: Special Issue on Semantic Integration, American Association for Artificial Intelligence" 2005, s. 45–58.

Mazurek M., *Architektura systemu wspomagania decyzji medycznych wykorzystująca technologię przetwarzania danych Big Data*, „Roczniki Kolegium Analiz Ekonomicznych" 2014, z. 35, s. 257–271.

Tekieh M.H., Raahemi B., *Importance of Data Mining in Healthcare: A Survey*, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Paris, August 25–28, 2015.

Zuccon G., Koopman B., Nguyen A., Vickers D., Butt L., *Exploiting medical hierarchies for concept-based information retrieval*, Proceedings of the Seventeenth Australasian Document Computing Symposium (ADCS '12), Dunedin, New Zealand, December 5–6, 2012, ACM, New York 2012.

Źródła sieciowe

Apache Flume, https://flume.apache.org/ (data odczytu: 20.11.2017).

CSIOZ, Klasyfikacje, https://www.csioz.gov.pl/interoperacyjnosc/klasyfikacje/ (data odczytu: 20.11.2017).

Hadoop and Mongo DB, https://www.mongodb.com/hadoop-and-mongodb (data odczytu: 20.11.2017).

International SNOMED CT Browser: http://browser.ihtsdotools.org/ (data odczytu: 20.11.2017).

Introducing JSON, https://www.json.org/ (data odczytu: 20.11.2017).

LOINC, https://loinc.org/ (data odczytu: 20.11.2017).

SNOMED, https://www.snomed.org/snomed-ct (data odczytu: 20.11.2017).

SNOMED CT Starter Guide, https://confluence.ihtsdotools.org/display/DOCSTART/SNOMED+CT+Starter+Guide (data odczytu: 20.11.2017).

* * *

Indexing the NoSQL Repository of Medical Records with Ontology Concepts

Abstract

Managing a huge amount of data coming from heterogenous sources with different schemes is a challenge when it comes to efficient querying data. Different systems use different terms to describe the same concepts. Traditional approaches based on the unification of data schema on input lack efficiency in processing high volumes of incoming data. The paper describes the system based on MongoDb schema-free database for medical records. The batch process is indexing data with equivalent concepts from SNOMED ontology. Aa s result, users and data mining tools can query databases solely with ontology concepts, and query results are in a tabular format, friendly for analytical tools.

Keywords: NoSQL, ontology, medical records database, MongoDB