UNIWERSYTET
EKONOMICZNY
W POZNANIU

Wydział Informatyki
i Gospodarki Elektronicznej

Włodzimierz Lewoniewski

Metoda porównywania i wzbogacania informacji w wielojęzycznych serwisach wiki na podstawie analizy ich jakości

The method of comparing and enriching information in multilingual wikis based on the analysis of their quality

Praca doktorska

Promotor: Prof. dr hab. Witold Abramowicz

Promotor pomocniczy: dr Krzysztof Węcel

Pracę przyjęto dnia:

podpis Promotora

Kierunek:

Specjalność:

Poznań 2018

Spis treści

Uwagi do pracy

Rozdział 1

Wstęp

1.1 Motywacja

W dzisiejszych czasach, aby podjąć prawidłowe decyzje gospodarcze, należy przeanalizować dużą ilość informacji i wiedzy. Informacja stała się towarem. Ilość i jakość informacji w dużym stopniu decydują o jakości decyzji w różnych gałęziach gospodarki. Z jednej strony menedżerowie dbają o dostęp do tak szerokiego zakresu informacji, jak to tylko możliwe. Z drugiej strony ważna jest także jakość informacji określona przez różne cechy (takie jak trafność, dokładność, jednoznaczność).

Informacja wysokiej jakości jest niezbędna do skutecznego działania i podejmowania decyzji w organizacji (Price i Shanks, 2016). Niedokładne i niekompletne informacje mogą negatywnie wpłynąć na przewagę konkurencyjną firmy (H. Xu i Koronios, 2005). Na przykład, w branży rekreacyjno-turystycznej istotna jest identyfikacja oraz ocena informacji środowiskowych do planowania biznesowego (de Freitas, 2003). Innym przykładem mogą być informacje o nowych terapiach, przydatne do podejmowania decyzji menedżerskich w szpitalach (Kidholm i in., 2015).

Internet umożliwia współdziałanie i wymianę informacji na skalę globalną. Zgodnie z Internet World Stats w czerwcu 2018 r. ponad połowa ludności świata korzystała z Internetu (**url20**).

Przydatne informacje można znaleźć zarówno w wyspecjalizowanych źródłach ekonomicznych, jak i w zasobach internetowych o charakterze ogólnym. Obecnie każdy może mieć swój wkład w rozwój wspólnej ludzkiej wiedzy w Internecie. Jednym z najlepszych przykładów takich repozytoriów online są serwisy internetowe typu wiki, w których treść można tworzyć i zmieniać z poziomu przeglądarki internetowej.

Najbardziej popularnym serwisem typu wiki jest Wikipedia[1]. Ta internetowa encyklopedia od ponad 15 lat istnieje jako ogólnie dostępny zasób, a każdy chętny może współtworzyć treści. Wikipedia stosunkowo szybko stała się ważnym źródłem informacji na całym świecie. Zawiera ponad 48 mln artykułów w około 300 różnych językach świata (Wikipedia Meta-Wiki, 2018b). Angielska (EN) wersja językowa jest największa i zawiera ponad 5,7 mln artykułów. Obecnie Wikipedia jest na piątym miejscu w rankingu najczęściej odwiedzanych stron w Internecie (**url21**), ustępując tylko Google[2], YouTube[3], Facebook[4] oraz Baidu[5]. W odróżnieniu od innych popularnych serwisów internetowych Wikipedia nie wyświetla reklam i utrzymuje się z darowizn od użytkowników w celu pokrycia kosztów bieżących serwisu.

Pomimo swojej popularności Wikipedia jest często krytykowana za niską jakość treści. Artykuły na określony temat mogą powstawać niezależnie w każdej wersji językowej. W związku z tym często jakość informacji o tym samym podmiocie czy wydarzeniu może się różnić w zależności od języka. Należy także zaznaczyć, iż temat opisany w jednej wersji językowej może zostać przetłumaczony na inne języki. Jednak stosunkowo mała liczba użytkowników posiadająca znajomość dwóch i więcej języków podejmuje taką inicjatywę poprzez przenoszenie treści pomiędzy różnymi wersjami językowymi.

Pomimo niekomercyjnego charakteru Wikipedii informacje zawarte w tej społecznej bazie wiedzy mogą wpływać na decyzje biznesowe i konsumenckie, ponieważ często strony tej encyklopedii pojawiają się jako pierwsze w wynikach wyszukiwania na różne tematy (**Lewandowski2011**). Strony Wikipedii o znanych osobach, firmach, produktach można znaleźć na pierwszych stronach w wynikach wyszukiwania Google, Bing[6], Yandex[7] oraz innych. Można się spodziewać, że osoby odwiedzające strony Wikipedii oraz jej twórcy są zainteresowani wysoką jakością treści w niej zawartej. Angielska wersja Wikipedii posiada najwięcej artykułów, jednak przedstawienie informacji w różnych językach jest szczególnie ważne dla użytkowników, którzy korzystają z wyszukiwarek w swoim ojczystym (nie angielskim) języku. Poza tym, niektóre tematy mogą być bardziej popularne na mniejszych obszarach, stąd większe jest prawdopodobieństwo znalezienia większej ilości informacji na dany temat w odpowiednich wersjach językowych (innych niż angielska). Są również tematy, które w ogóle nie są opisane w angielskiej

[1] http://www.wikipedia.org
[2] http://www.google.com
[3] http://www.youtube.com
[4] http://www.facebook.com
[5] http://www.baidu.com
[6] http://www.bing.com
[7] http://www.yandex.ru

wersji Wikipedii, a pojawiają się w mniej rozwiniętych wersjach językowych Wikipedii (Bao i in., 2012; Callahan i Herring, 2011).

W Wikipedii istnieje system oceny jakości artykułów, jednak konkretna wersja językowa może używać własnych standardów oraz skali ocen. Praktycznie każda wersja językowa ma specjalne wyróżnienie dla artykułów o najlepszej jakości. W angielskiej wersji takie artykuły nazywają się „Featured Articles" (FA), w polskiej – „Artykuły na Medal". Takie najlepsze artykuły powinny spełniać określone kryteria jakości związane m.in. z dokładnością, neutralnością, kompletnością i stylem. Przewidziane jest też wyróżnienie dla artykułów wysokiej jakości, które nie spełniają wszystkich kryteriów FA - „Good Articles" (GA), w języku polskim to „Dobre artykuły". Liczba artykułów FA i GA jest stosunkowo bardzo mała w każdej wersji językowej – zazwyczaj mniej niż 1% całkowitej liczby artykułów. W niektórych wersjach językowych Wikipedii istnieją również oceny dla artykułów gorszej jakości. W angielskiej Wikipedii dodatkowo stosowane są następujące klasy jakości: A-class, B-class, C-class, Start, Stub. Każda z tych ocen może pokazywać stopień rozwoju artykułu. Jednak nie wszystkie wersje językowe stosują tak rozwiniętą skalę ocen. Np. niemiecka Wikipedia stosuje tylko dwie oceny dla artykułów najwyższej jakości (odpowiedniki FA i GA), białoruska ma 3 oceny (FA, GA, Stub). Rosyjska Wikipedia posiada 7 ocen, jednak nie wszystkie z ich mają odpowiedniki w wersji angielskiej (FA, GA, SA, I, II, III, IV). Dodatkowo należy wspomnieć o dużej liczbie nieocenionych artykułów. Np. w wersji niemieckiej i polskiej udział takich artykułów wynosi ponad 99%, w ukraińskiej takich artykułów jest ponad 97%, a w wersji rosyjskiej – ponad 85%.

Obecnie istnieje szereg podejść, które umożliwiają z różnym stopniem precyzji automatycznie ocenić jakość artykułów w ramach określonej wersji językowej. Takie badania skupiają się głównie na wersji angielskiej. Jednym z pierwszych badań w tym kierunku jest analiza objętości treści artykułów (Stvilia, Twidale, Smith i Gasser, 2005a). Takie proste miary, jak liczba słów, mogą pomóc w ocenie jakości artykułów z Wikipedii (Blumenstock, 2008b). Najlepsze artykuły korzystają również z większej liczby referencji i zawierających więcej sekcji (Warncke-wang, Cosley i Riedl, 2013). Dodatkowo do analizy mogą być brane pod uwagę specjalne szablony o lukach w jakości, dotyczących wiarygodności, stylu pisania, czy struktury (Anderka, 2013).

W przypadkach, kiedy objętość treści w artykułach jest podobna, lepszy artykuł będzie mieć więcej informacji faktycznej (Lex i in., 2012). Styl i różnorodność wykorzystanych słów również wpływa na jakość artykułu (Lipka i Stein, 2010; Y. Xu i Luo, 2011).

Ocena jakości artykułów Wikipedii może opierać się nie tylko na miarach związanych z treścią. Inne badania pokazują, jak miary związane z reputacją użytkowników i siecią autorów, stanem artykułu, zewnętrznym wsparciem faktycznym i innymi wskaźnikami mogą pomóc w określeniu jakości artykułu (Velázquez, Cagnina i Errecalde, 2017; G. Wu, Harrigan i Cunningham, 2011).

Wiele z tych badań rozwiązuje problem automatycznej oceny artykułów jako zadanie klasyfikacji - wszystkie oceny są podzielone na dwie grupy: artykuły kompletne i niekompletne (Lex i in., 2012; Warncke-wang i in., 2013). Grupa kompletnych artykułów składa się z artykułów ocenionych jako FA i GA. Pozostałe oceny niższej jakości są zawarte w grupie niekompletnych artykułów. Do budowania modeli są wykorzystywane różne miary artykuły Wikipedii, które są zmiennymi niezależnymi, natomiast jakość - stochastyczna zmienna zależna (Warncke-wang i in., 2013). Badania wykazały, że istnieją różnice między modelami jakości w poszczególnych wersjach językowych Wikipedii przy użyciu tego samego zbioru zmiennych niezależnych (miar). Najczęściej do budowy takich modeli stosowane są algorytmy eksploracji danych, a w szczególności Random Forest (losowy las), który pokazał największą precyzję w klasyfikacji (Warncke-wang i in., 2013).

Używanie miary stochastycznej do oceny jakości artykuły zazwyczaj daje wysoki poziom precyzji w modelach klasyfikacji (ponad 95% w różnych wersjach języków Wikipedii), jednak to podejście ma pewne wady i ograniczenia. Jeżeli artykuły należą do tej samej grupy (np. niekompletne), to nie jest możliwe porównanie ich jakości pomiędzy sobą.

W niektórych pracach jakość w modelach rozpatrywana jest również jako zmienna kategorialna (Dang i Ignat, 2016b; Warncke-wang i in., 2013). Jednak poziom precyzji w takich modelach jest znacznie mniejszy niż w przypadku używania stochastycznej zmiennej zależnej – około 60%. Dodatkowo w takim podejściu porównanie jakości artykułów w różnych językach będzie poważnym wyzwaniem w związku z różnicami w systemach klasyfikacji ocen w poszczególnych wersjach językowych Wikipedii.

Znacznie bardziej użytecznym rozwiązaniem może być stosowanie zmiennej ciągłej do oceny jakości artykułów. Używając różnych zmiennych ilościowych artykułów (takich jak długość tekstu, liczba obrazów, referencji, sekcji itp.) w celu obliczenia tzw. relatywnej jakości tego samego artykułu w różnych wersjach językowych Wikipedii (np. w skali od 0 do 100). Miary do analizy jakości stron wiki mogą być ekstrahowane w różny sposób.

Jednym z ważniejszych elementów stron wiki są referencje. Większość badań zazwyczaj skupia się na liczeniu referencji na stronie oraz używa tej liczby do tworzenia innych (pochodnych) miar (np. ref./długość). Badanie jakości oraz podobieństwa referencji pomiędzy różnymi wersjami językowymi strony wiki na określony temat może polepszyć istniejące modele jakości.

W artykułach Wikipedii może być umieszczona specjalna wyróżniona ramka, która w przejrzysty sposób prezentuje najważniejsze dane. Ta ramka ma nazwę „infoboks" i zazwyczaj umieszczana jest w prawym górnym rogu artykułu. Każdy element infoboksu to para parametr i jego wartość (infoboks będzie rozumiany jako lista elementów infoboksu). W zależności od tematu infoboks może zawierać pewien zestaw dopuszczalnych parametrów. Dane z infoboksów mogą służyć nie tylko do szybkiego zapoznania się z tematem przez czytelnika Wikipedii, ale również do wzbogacenia innych popularnych baz danych, takich jak DBpedia. Z tego powodu szczególnie istotna jest weryfikacja jakości wprowadzanych przez użytkowników danych.

Jakość infoboksów jest znacznie mniej rozwiniętym tematem w literaturze naukowej. Istniejące badania często analizują jakość baz danych utworzonych na podstawie informacji zamieszczanych w infoboksach. Dobrym przykładem takich baz jest DBpedia, która dodatkowo zawiera wiele linków do innych zestawów danych takich, jak Freebase, OpenCyc (Färber, Bartscherer, Menne i Rettinger, 2016).

DBpedia może być stosowana nie tylko do unifikacji oraz lepszej organizacji danych pochodzących z różnych wersji językowych Wikipedii. Ta semantyczna baza wiedzy potrafi skutecznie wspomagać aplikacje do rozpoznania jednostek w tekście, wzbogacać systemy udzielania odpowiedzi na pytania i szukania słów kluczowych, wprowadzać informacje kontekstowe dla rekordów bibliograficznych i archiwalnych (Lehmann i in., 2015).

Korzystając z zestawu ogólnych wzorców testu jakości danych, można analizować różne kwestie związanych z jakością danych (Kontokostas i in., 2014). Stosując specjalne metody, można analizować zwięzłość, spójność, trafność składni (syntaktyka) i dokładność semantyczną danych, zawartych w DBpedii (Mihindukulasooriya, Rico, García-Castro i Gómez-Pérez, 2015). Analiza jakości danych w tej semantycznej bazie wiedzy jest możliwa również bez wykorzystania ontologii (Jang, Megawati, Choi i Yi, 2015). Istnieją badania związane z fuzją danych z różnych wersji językowych DBpedii (Tacchini, Schultz i Bizer, 2009). Jednak zdecydowana większość prac nie uwzględnia różnych aspektów jakości infoboksów i stron wiki, z których pochodzą te dane.

Jednym z ważniejszych miar jakości elementów infoboksu może być popularność określonego tematu w danej wersji językowej. Na przykład popularność pisarza w określonej wersji

językowej Wikipedii zależy od tego, czy ten pisarz jest związany z danym językiem (Hube, Fischer, Jäschke, Lauer i Thomsen, 2017).

Ocena jakości infoboksów dla poszczególnych tematów pozwoli na wybranie tych wersji językowych, gdzie umieszczone dane mają najlepszą jakość. To w konsekwencji może pomóc w poprawieniu jakości i wzbogaceniu innych wersji językowych Wikipedii.

Podsumowując, analiza stosowanych obecnie podejść do oceny jakości informacji w serwisach typu wiki pokazuje, że prowadzenie dalszych badań w celu opracowania nowych metod wydaje się uzasadnione i potrzebne. Rezultaty otrzymane w wyniku zastosowania takich metod mogą pozwolić na bardziej precyzyjną ocenę jakości informacji w serwisach typu wiki w różnych językach i tym samym przyczynić się do poprawy jakości informacji.

Przedstawiona praca dotyczy zagadnienia oceny jakości informacji zawartych na stronach wiki poprzez opracowanie autorskiej metody do porównywania i wzbogacenia informacji w wielojęzycznych serwisach wiki na podstawie analizy ich jakości. Opracowana metoda zostanie następnie poddana ewaluacji na podstawie rzeczywistych danych, pochodzących z 5 wersji językowych Wikipedii: angielska (EN), rosyjska (RU), polska (PL), ukraińska (UK), białoruska (BE).

Zgodnie z klasyfikacją dziedzin i dyscyplin naukowych w Polsce prezentowana rozprawa plasuje się w obszarze nauk społecznych, w dziedzinie nauk ekonomicznych, w dyscyplinie ekonomia (specjalność: informatyka ekonomiczna). Zgodnie z klasyfikacją Journal of Economic Literature (JEL) praca porusza następujące zagadnienia: C55 (Large Data Sets: Modeling and Analysis), D8 (metodologia gromadzenia oraz estymacji danych, ang. Data Collection and Data Estimation Methodology), L15 (Information and Product Quality; Standardization and Compatibility), L86 (Information and Internet Services; Computer Software)

1.2 Cel badawczy i teza pracy

Celem proponowanej rozprawy doktorskiej jest **opracowanie metody porównywania oraz wzbogacania informacji w wielojęzycznych serwisach wiki na podstawie analizy ich jakości na przykładzie Wikipedii**.

Proponowana metoda różni się od stosowanych dotychczas podejść pod kilkoma względami. Po pierwsze, w dotychczasowych pracach analiza jakości przeprowadzona była głównie w ramach jednej wersji językowej – najczęściej dla artykułów angielskiej Wikipedii. Niektóre zmienne (miary), które można brać pod uwagę przy budowaniu modeli jakości artykułów, są

zależne od języka w którym te artykuły są napisane. Dotyczy to m.in. miar lingwistycznych. Po drugie, brak badań, które w sposób automatyczny pozwalałyby mierzyć oraz porównywać jakość wybranego artykułu Wikipedii w różnych wersjach językowych. To jest związane m.in. z różnicami w systemach ocen stosowanych w każdej wersji językowej Wikipedii. Po trzecie, dotychczasowe prace skupiają się głównie na jakości całego artykułu, a nie poszczególnych jego ważnych elementów, takich jak infoboksy. Wstępne badania pokazują, że nie zawsze artykuł posiadający najwyższą ocenę spośród innych języków ma również infoboks z danymi o najlepszej jakości w danej wersji językowej.

Ponadto większość badań stosuje określony zbiór miar do budowy modeli jakości artykułów Wikipedii. Dobór niektórych z tych miar jest zależny od języka, inne miary zależą od źródła danych czy też sposobu ekstrakcji. Dodatkowym czynnikiem jest rozwój technologii serwisów wiki, który umożliwia opracowanie nowych miar. Oznacza to, że zebranie i łączenie wielu miar na podstawie literatury oraz własnych eksperymentów może pozwolić na bardziej wszechstronne i wiarygodne podejście do analizy jakości artykułów Wikipedii w różnych językach.

Dodatkową kwestią jest ciągła aktualizacja i pojawianie się nowych stron, na podstawie których budowane są modele jakości artykułów Wikipedii. Czynnik czasowy jest istotny nie tylko z powodu zmieniającej się liczby artykułów, ale również z powodu nieustannej aktualizacji zasad i reguł oceniania artykułów przez społeczność użytkowników Wikipedii w każdej wersji językowej. W związku z tym, artykuły, które wcześniej były wyróżnione najwyższą oceną, po pewnym czasie mogą już nie spełniać wymaganych kryteriów i stracić wyróżnienie.

Dodatkowo zdefiniowano cele pomocnicze, które przyczyniają się do realizacji celu głównego:

- Opracowanie metody automatycznej oceny jakości strony wiki w różnych językach z wykorzystaniem odpowiednich miar. W ramach rozprawy przeanalizowane zostały miary spotykane w literaturze i uwzględniane w istniejących rozwiązaniach, zaproponowane zostaną nowe miary, dotychczas nie brane pod uwagę. Dodatkowo została opracowana typologia miar oraz wymiary jakości stron wiki.
- Opracowanie metody porównania jakości infoboksu z jakością strony wiki. Metoda ma na celu znalezienie miar jakości strony wiki oraz infoboksów, które są mocno skorelowane między sobą. Uwzględnione zostały miary dotyczące m.in takich wymiarów jak kompletność, wiarygodność i aktualność.

- Opracowanie metody identyfikacji infoboksów oraz parametrów w nim umieszczonych o najwyższej jakości spośród odpowiedników strony wiki w różnych wersjach językowych. W pracy została przedstawiona metoda, która pozwoli na analizę jakości infoboksów oraz poszczególnych jego parametrów w każdej wersji językowej wiki w celu identyfikacji elementów o najwyższej jakości.
- Opracowanie metody wzbogacenia infoboksów pomiędzy wersjami językowymi wiki z wykorzystaniem semantycznej reprezentacji elementów tych infoboksów. W ramach rozprawy została opracowana metoda, która pozwala na przenoszenie wybranych elementów infoboksów o najlepszej jakości stron wiki określonej wersji językowej do odpowiedników tej strony w innych językach.
- Opracowanie metody tworzenia nowej strony w określonej wersji językowej z wybranymi elementami infoboksu o najwyższej jakości z innych wersji językowych wiki.

Metoda ma na celu tworzenie nowych stron wiki w określonych wersjach językowych, do których zostaną przeniesione infoboksy wraz z elementami o najwyższej jakości z odpowiedników tej strony wiki w innych językach.

Przyjęta w rozprawie teza brzmi następująco:

Metoda oceny jakości strony wykorzystująca semantyczne powiązania z innymi wersjami językowymi oraz uwzględniająca popyt na informację pozwala na porównanie i wzbogacenie informacji w serwisach wiki.

1.3 Źródła informacji i metody badawcze

Wyniki analizy literatury zostały przedstawione w czterech pierwszych rozdziałach pracy. Analiza obejmuje ponad sto publikacji naukowych (przede wszystkim artykułów i monografii angielskojęzycznych, znajdujących się w zasobach baz ACM Digital Library, SpringerLink, ProQuest, IEEE Xplore, Scopus) z takich dziedzin, jak ekonomia i informatyka. Uzupełnienie stanowi przegląd kilkudziesięciu źródeł internetowych. Wynikiem analizy literaturowej w rozprawie było: objaśnienie problemu badawczego, charakterystyka podejść stosowanych do analizy jakości stron wiki oraz infoboksów, opracowanie typologii miar jakości artykułów i infoboksów dotyczących m.in. takich wymiarów jak kompletność, wiarygodność, aktualność. Przeprowadzone analizy umożliwiły opracowanie koncepcji metody, która na podstawie oceny jakości strony i wykorzystująca semantyczne powiązania z innymi wersjami językowymi oraz uwzględniająca

popyt na informację pozwala na porównanie oraz wzbogacenie informacji w serwisach wiki. Badania empiryczne nad metodą zostały podzielone na następujące etapy:

- Przygotowanie zbioru danych testowych obejmujących informację o stronach Wikipedii w 5 wersjach językowych oraz dodatkowe dane dotyczące popytu na te strony.
- Przygotowanie danych do analizy referencji z otwartych źródeł.
- Ekstrakcja miar stron wiki oraz infoboksów z danych testowych, w tym dotycząca popytu na te strony.
- Dobranie algorytmów klasyfikacyjnych do automatycznej predykcji jakości stron wiki na podstawie ekstrahowanych miar.
- Identyfikacja istotnych miar do analizy jakości stron przy użyciu wybranego algorytmu klasyfikacyjnego.
- Przygotowania danych do przeprowadzenia analizy jakości danych z infoboksu.
- Przeprowadzenie analizy współzależności miar stron wiki oraz miar infoboksów.
- Przeprowadzenie eksperymentów mających na celu określenie dokładności predykcji.
- Zainstalowanie serwisu wiki do testowania metody wzbogacenia stron.
- Przeprowadzenie eksperymentów związanych z przenoszeniem elementów infoboksów strony w określonej wersji językowej do odpowiednika tej strony w innych językach.
- Przeprowadzenie eksperymentów związanych z utworzeniem nowej strony w określonej wersji językowej dla przenoszenia elementów infoboksu z innej wersji językowej.
- Ewaluacja wyników metodą ekspercką.

Prace badawcze prowadzone są zgodnie z paradygmatem projektowania (ang. Design Science) Hevnera (Hevner, 2004), który rozwiązuje problemy badawcze przez projektowanie nowych rezultatów (ang. *artifacts*). Poszczególne rezultaty opracowywane są zgodnie z modelem Boehma (model spiralny) (Boehm, 1988), w którym każda ze spirali składa się z następujących etapów:

- a) definiowanie celów i rozpoznawanie wymagań oraz zagrożeń;
- b) analiza istniejących metod związanych z określonymi celami szczegółowymi;
- c) opracowanie nowych rozwiązań, zgodnych z przyjętymi założeniami;
- d) ocena postępów prac polegająca na przeprowadzeniu eksperymentów, analizie ich wyników oraz na planowaniu dalszych działań (wprowadzeniu modyfikacji i poprawek w celu poszukiwaniu najlepszego możliwego rozwiązania) lub zakończenie badań.

1.4 Struktura rozprawy

Praca składa się z dwunastu rozdziałów, które pogrupowane są na dwie części. Pierwsza część (nieoryginalna) poświęcona jest analizie literatury. W części drugiej (oryginalnej) przedstawiono opracowane w trakcie badań artefakty. Strukturę pracy dopełnia wstęp oraz podsumowanie.

Rozdział pierwszy przedstawia motywację badań, cel badawczy wraz z tezą pracy, źródła oraz metody badawcze.

Rozdział drugi poświęcony analizie znaczenia informacji w gospodarce. Również ten rozdział zawiera analizę literatury w zakresie jakości informacji: zawiera on podstawowe definicje związane z jakością danych, jakością oraz wymiarami jakości informacji.

W ramach rozdziału trzeciego omówione zostały serwisy wiki, które umożliwiają współtworzenie treści za pośrednictwem Internetu oraz przeglądarki. W szczególności została opisana otwarta encyklopedia Wikipedia, która jest najbardziej znanym przykładem serwisów wiki. Dodatkowo zostały opisane infoboksy, które często umieszczane są w widocznych miejscach artykułów Wikipedii oraz które są istotne z punku widzenia niniejszej rozprawy. Dane z infoboksów mogą być przydatne do automatycznego wzbogacenia różnych semantycznych baz wiedzy. To takich baz należy również DBpedia, która została opisana w ramach trzeciego rozdziału.

Rozdział czwarty poświęcony jest analizie literatury w zakresie metod do automatycznego określenia jakości artykułów Wikipedii. W tym rozdziale m.in. zostały omówione wymiary oraz problemy jakości serwisów wiki.

Rozdział piąty opisuje miary oraz wymiary jakości artykułów Wikipedii. Tutaj została przedstawiona autorska typologia miar jakości z uwzględnieniem jej wymiarów. Przedstawione zostaną również algorytmy wyznaczania miar jakości artykułów Wikipedii, opis metod ich ekstrakcji wraz ze źródłami.

W rozdziale szóstym zostaną zweryfikowane istniejące podejścia do predykcji jakości oraz zaproponowane i przetestowane nowe sposoby mierzenia jakości. W szczególności zostaną opisane miary dotyczące popytu na informację, metodę analizy podobieństwa referencji literaturowych, miary SEO oraz inne.

W ramach rozdziału siódmego zostaną przedstawione miary oraz wymiary jakości dotyczące infoboksów. Dodatkowo zostały pokazane zasady oraz wyniki obliczenia wybranych miar dla niektórych tematów Wikipedii.

Rozdział ósmy poświęcony budowaniu modelu jakości infoboksów, który wykorzystuje opisane wcześniej miary jakości. Została również przeprowadzona analiza współzależności miar jakości infoboksów oraz artykułów.

Rozdział dziewiąty opisuje autorską metodę porównywania informacji w różnych językach na podstawie analizy jakości. Przy tym wykorzystywane są dane na określone tematy z różnych wersji językowych Wikipedii.

W ramach rozdziału dziesiątego została przedstawiona autorska metoda wzbogacenia informacji na podstawie analizy ich jakości w różnych wersjach językowych. Zostaną pokazane wyniki działania tej metody na konkretnych przykładach z Wikipedii.

Jedenasty rozdział poświęcony jest ewaluacji autorskiej metody przedstawionej w rozprawie. Metoda została zweryfikowana na podstawie zgodności z ocenami ekspertów, którzy przeprowadzili ocenę danych konkretnych infoboksów.

W ostatnim rozdziale zostały przedstawione wnioski wraz z oceną osiągniętych celów badawczych oraz weryfikację tezy. Dodatkowo zostały opisane możliwe kierunki dalszych badań.

Rozdział 2

Jakość danych i informacji

Celem niniejszego rozdziału jest przegląd istniejących podejść do definiowania pojęcia informacji oraz jej jakości. W szczególności zostaną przeanalizowane proponowane miary jakości informacji. Analiza stanu wiedzy zostanie przeprowadzona zgodnie z metodyką Webstera (Webster i Watson, 2002).

Obszerne materiału tego rozdziału zostały opracowane na podstawie wcześniejszych badań (Lewoniewski, Węcel i Abramowicz, 2015).

2.1 Wprowadzenie

Relacje pomiędzy danymi, informacją i wiedzą są trudno definiowalne (Abramowicz, 2008). Dane opisują stan i własności rzeczy, osób, zjawisk oraz pojęć abstrakcyjnych. Dane mogą przyjmować różne postaci: mowa, tekst, rysunki, sygnały. Informacja – to dane zawarte w komunikacie, zinterpretowane przez odbiorcę, mające dla niego znaczenie (Swoboda, 2015). Innymi słowy, dane stają się informacją po zinterpretowaniu ich przez ludzi (Abramowicz, 2008). Jeżeli te informacje zostały zrozumiane taki sposób, że wyjaśniają czy pozwalają zrozumieć coś, mówimy wtedy o pojęciu wiedzy (Jennex i Bartczak, 2013). Następnym poziomem abstrakcji jest mądrość - umiejętność wykorzystywania wiedzy. Relacje pomiędzy przedstawionymi pojęciami przedstawia tzw. piramida wiedzy (patrz rys. 2.1), która pokazuję te pojęcia w postaci piramidy.

Rysunek 2.1. Piramida wiedzy.

Źródło: (Abramowicz, 2008)

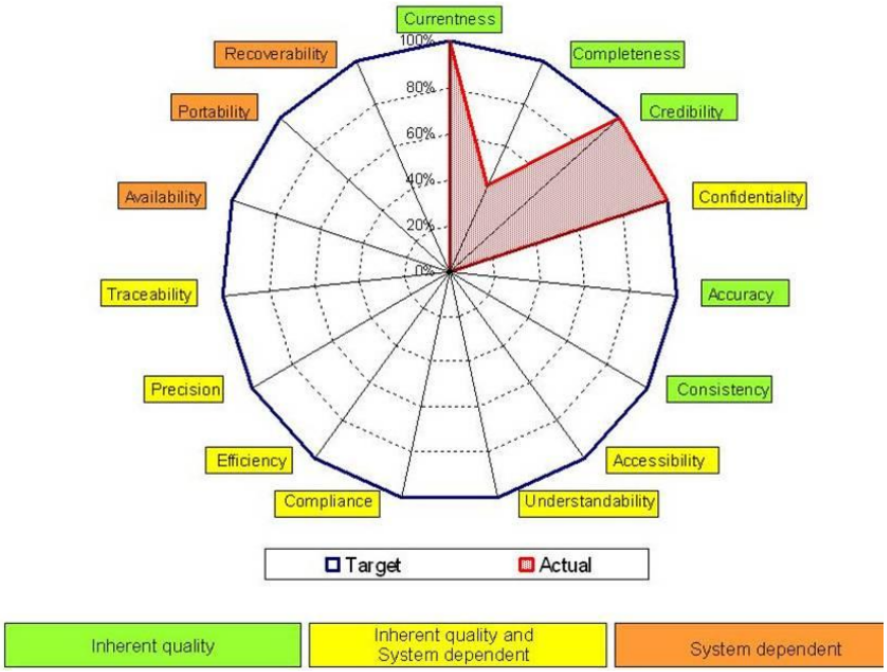
2.2 Jakość danych

Pojęcie jakości było definiowane już w czasach starożytnych. Platon wnioskował, że jakość jest sądem wartościującym wyrażonym przez użytkownika. Chiński filozof Lao Tse potraktował jakość jako doskonałość, do której trzeba konsekwentnie dążyć. W latach 1970Ќolman wyróżnił pięć grup kryteriów jakościowych: przydatność, poprawność, użyteczność, doznaniowość, opłacalność. Każda grupa to stopień spełnienia odpowiednich wymagań. Później Garvin wskazał na siedem kategorii definicji jakości (Garvin, 1984), które są najczęściej wskazywane w najnowszych badaniach: ogólne, wielowymiarowe, strategiczne oraz związane z produktem, produkcją, użytkownikiem lub z tworzeniem wartości.

Przy definiowaniu jakości wyróżniamy podejścia deskryptywne i wartościujące. W przypadku podejścia deskryptywnego mamy do czynienia z pojęciem abstrakcyjnym, kiedy jakość oznacza wzór, do którego należy dążyć. W takim podejściu trudno wprowadzić miary jakości. Inaczej jest w przypadku podejścia wartościującego, gdzie jakość rozumiana jest jako poziom spełnienia wymagań dotyczących produktu. To jest szczególnie aktualne w obecnych czasach, kiedy nabywca produktu indywidualnie dokonuje oceny jakości (Swoboda, 2015).

Bardzo ważną cechą jakości danych jest zdolność do spełnienia wymagań użytkownika końcowego (Benson, 2008). Dane wysokiej jakości muszą spełniać określone kryteria, które są definiowane różnie w zależności od standardów.

Jednym z takich standardów jakości jest ISO 8000. Zgodnie z tą normą w celu zapewnienia jakości danych należy brać pod uwagę szereg kryteriów, wliczając w to takie cechy danych, jak: dokładność, źródło, kompletność, cel (przeznaczenie) danych, metoda pomiaru lub osza-



Rysunek 2.2. Jakość danych w oparciu o ISO 25012.

Źródło: (Natale, 2011)

cowania (Grantner, 2007). Ten standard też opisuje działania na rzecz sprawdzania wskaźników jakości danych (ISO/TS, 2011), na przykład:

- Sprawdzenie zgodności danych ręcznie lub przy użyciu narzędzia
- Statystyczna analiza wskaźników jakości danych
- Mogą być brane pod uwagę otwarte słowniki techniczne zgodnie z ISO/TS 22745-30:2009

Innym kluczowym standardem jakości danych jest ISO/IEC 25012 (Aljumaili, Karim i Tretten, 2016; Natale, 2011). Rys. 2.2 przedstawia kryteria jakości danych w oparciu o ten standard.

Badania dotyczące oceny jakości danych wskazują również na procedury, które mogą pomóc tę jakość poprawić. Na przykład integracja danych poprawia użyteczność danych poprzez poprawę spójności, kompletności, dostępności oraz innych atrybutów danych (Madnick, Wang, Lee i Zhu, 2009).

Aktualność danych jest szczególnie ważna w zarządzaniu jakością danych. Heinrich i Klier zaproponowali dla mierzenia aktualności danych wykorzystać metrykę PBCM (probability-based currency metric). W ramach tej metryki aktualność jest interpretowana jako prawdopodobieństwo, że wartość atrybutu jest wciąż taka sama jak wartość tego atrybutu w świecie rzeczywistym w chwili oceny (Heinrich i Klier, 2015).

2.3 Jakość informacji

Ogólnie w literaturze nie ma zgodności na temat rozróżnienia pojęć jakość danych i jakość informacji. Mandrick i inni zaznaczyli, że istnieje tendencja do używania pojęcia jakości danych w odniesieniu do zagadnień technicznych (np. integracja danych), natomiast jakość informacji – do zagadnień nietechnicznych, np. relewancja dla konkretnego odbiorcy (Madnick i in., 2009). Z drugiej strony, pojęcie jakości informacji, podobnie jak pojęcie jakości danych, można scharakteryzować jako rozbieżność pomiędzy wizją świata dostarczoną przez system informacyjny i prawdziwego stanu świata (Parssian, Sarkar i Jacob, 2004).

Odmienność sposobów definiowania i traktowania informacji przez badaczy nie pozwala znaleźć powszechne akceptowanej definicji jego jakości. Zatem mamy różnorodne koncepcje określenia pojęcia informacji w zależności od przyjmowanych założeń wyjściowych. Na przykład, w ujęciu ekonomicznym informacja może być rozważana jako towar lub zasób, kapitał lub część procesu komunikacji. Jakość tego produktu jest rozumiana jako pojęcie subiektywne: każdy użytkownik może inaczej oceniać otrzymaną informację. Ocena jakości może być zależna od wiedzy, doświadczenia i popytu informacyjnego konkretnego odbiorcy. Na przykład bardzo dokładna informacja z obszaru zainteresowania jednej osoby będzie mieć mniejszą jakość dla osoby spoza grupy docelowej tego komunikatu.

Metody i kryteria oceny jakości informacji dobierane są w zależności od rodzaju informacji: ekonomiczna, medyczna, techniczna itd. Na przykład dla stron internetowych związanych z ochroną zdrowia są przyjęte specjalne kryteria oceny jakości przez Komisję Wspólnot Europejskich. W tym przypadku jakość strony internetowej mierzona jest na podstawie następujących atrybutów: przejrzystość, uczciwość, dostępność, wiarygodność, aktualność, ochrona danych osobowych, odpowiedzialność (Commission of the European Communities, 2002).

W aspekcie ekonomicznym informacja oceniana jest z punktu widzenia przydatności dla osoby, która podejmuje decyzje. W tym przypadku jakość informacji głównie jest mierzona na podstawie następujących cech: aktualność, dokładność, zrozumiałość, zwięzłość, trafność. W praktyce rzadko się zdarza, aby informacje posiadały wszystkie atrybuty jednocześnie. Najczęściej to jest uzależnione od źródeł informacji i od umiejętności ich gromadzenia (Bartosik-Purgat, Mruk i Schroeder, 2012). W epoce przeładowania informacyjnego szczególną rolę odgrywa zwięzłość, która pozwala odbiorcy uniknąć zbędnej straty czasu i jak najszybciej otrzymać istotne informacje. W tym sensie jednym z dogodnych źródeł informacji może być encyklopedia.

Przy ocenie ilościowej jakości informacji w Internecie można posługiwać się różnymi metodami. Jedna z nich to metoda statystyczna, która wyciąga wnioski z analiz liczbowych różnorodnych danych (Bednarek-Michalska, 2007). Takimi danymi ilościowymi mogą być metadane. Do metadanych dokumentu można odnieść atrybuty, mówiące o jego treści, strukturze i źródle pochodzenia. Na przykład: rozmiar dokumentu (liczba znaków, ilustracji i itp.), powiązania z innymi dokumentami (Abramowicz, 2008).

Ocena jakości informacji może być rozpatrywana również w ramach koncepcji infologicznej. W ogólnym ujęciu, opartym na propozycji Sundgrena, opis obiektu *O* może być przedstawiony w postaci układu (Sundgren, 1973):

$$K : \langle O, X, x, t \rangle \tag{2.1}$$

gdzie: *O* – obiekt należący do analizowanej rzeczywistości R; *X* – cecha, ze względu na którą obserwator U analizuje obiekt O, *x* – wartość cechy X, *t* – czas, w którym obiekt O przyjmuje wartość x cechy X.

Do komunikatu K w układzie (równanie 2.1) można dopisać wektor dodatkowych charakterystyk związanych z obiektem *O*, atrybutem *X* i jego wartością x oraz czasem *t* (Stefanowicz, 2010). W tym przypadku możemy zastosować metodę ekspercką (jakościowo-heurystyczną), kiedy analizowane są cechy treściowe informacji.

Swoboda (2015) w swojej pracy dla oceny jakości informacji proponuje stosować formularz, który jest konstruowany przez oceniającego. Wstępnie osoba musi ustalić kryteria (mierzalne i niemierzalne) oraz określić stopień ich ważności. Pomiar dla kryteriów niemierzalnych może polegać na określeniu stopnia spełnienia wymagań. Najprostszym sposobem jest stosowanie podejścia binarnego (spełnia lub nie spełnia). Grudzień przy ocenianiu jakości informacji o procesach w systemach zarządzania proponuje stosowanie arkuszu badawczego cech procesu oraz arkusza, który gromadzi dane na podstawie opisów jakościowych poszczególnych atrybutów informacji (Grudzień, 2012).

Madnick i in. (2009) na podstawie analizy różnych znanych podejść oceny jakości danych zaproponowali klasyfikator, który pozwala rozróżnić modele oceny jakości danych pod kątem dwóch wymiarów: temat i metoda. Warto zwrócić uwagę na słowa kluczowe odnoszące się do wymiaru „metoda". Są to m.in.: badanie działania, studium przypadku, eksploracja danych, sztuczna Inteligencja, ilościowa, eksperymentalna, modelowanie matematyczne, ekonometria, analiza statystyczna. Badanie działania opiera się na metodzie empirycznej i interpretacyjnej.

Prace w tym kierunku badają zachowanie doświadczonych praktyków, którzy zajmują się rozwiązaniem problemów jakości danych (Lee, 2003). W przypadku sztucznej inteligencji, techniki automatycznego wnioskowania i reprezentacja wiedzy mogą pomóc w poprawianiu wymiaru przyswajalności i spójności jakości danych (Madnick i Zhu, 2006).

Metody oceny jakości informacji można podzielić również na obiektywne i subiektywne. W produkcyjnych systemach informacyjnych dane surowych oraz składowe są oceniane przez obiektywną metodę oceny jakości informacji. Natomiast jakość produktów informacyjnych, które są skierowane do klienta, jest szacowana na podstawie subiektywnej metody oceny jakości (Ge i Helfert, 2008). Aby dokładniej wyjaśnić różnice między obiektywną i subiektywną oceną, trzeba porównać je pod kątem następujących aspektów: narzędzia, obiekt pomiaru, kryteria, proces, wynik oceny i przechowywanie danych. W przypadku obiektywnego oceniania są wykorzystywane reguły lub wzorce, które mogą w sposób automatyczny ocenić jakość danych w bazie. Dla subiektywnej oceny trzeba poprosić każdego uczestnika badania o oszacowanie, czy dany produkt informacyjny jest dla niego przydatny (ewentualnie w jakim stopniu). Różnica w tych podejściach może doprowadzić do sytuacji, kiedy obiektywna metoda oceny będzie nadawała wysoką ocenę danym, które nie są przydatne dla użytkownika. Miary obiektywne są bowiem najczęściej niezależne od użytkownika. Na przykład bardzo szczegółowa informacja może być trudna do zrozumienia. Dla obiektywizacji oceny informacji Ge i Helfert (2008) proponują strategię 5 kroków: sporządzenie specyfikacji, identyfikacja problemów jakości informacji, powiązanie problemów jakości informacji z jej wymiarami, ocena jakości informacji, generowanie raportu. W subiektywnej ocenie jakości często stosowane jest podejście statystyczne dla identyfikacji najważniejszych czynników wpływających na jakość informacji (H. Xu, 2015). Matematyczne modelowanie w ocenie jakości informacji również znajduje zastosowanie.

Mierzenie jakości informacji zazwyczaj wykonywane jest w odniesieniu do jakiejś równowagi społecznej dotyczącej tego, co stanowi dobrą jakość informacji w danym kontekście aktywności (Stvilia, Gasser, Twidale i Smith, 2007). Sam kontekst składa się z dwóch głównych elementów: kultury (języka, normy) i struktur społeczno-technicznych (w tym stosunków gospodarczych i standardów).

Podobnie jak w przypadku danych, dla dokonania oceny jakości informacji trzeba brać pod uwagę jakość wszystkich jej atrybutów (Abramowicz, 2008). Atrybuty informacji mogą być tożsame z wymiarami jakości informacji (Singh, Singh, Park i Lee, 2009).

Istotność każdego atrybutu informacji zależy od bieżących potrzeb użytkownika informacji. Takie zjawisko można obserwować nawet w przypadku różnych informacji dotyczących tego samego obszaru. Na przykład badania dotyczące przepływu informacji o katastrofach wykryły, że przy ocenie jakości informacji przy każdej katastrofie różnie będzie postrzegany poziom istotności takich atrybutów jak: dostępność, aktualność, precyzja, dokładność (Singh i in., 2009).

Atrybuty informacji można podzielić na informacyjne (w tym prawdziwość, aktualność, wiarygodność, użyteczność, przyswajalność i inne) i techniczne (w tym objętość, medium informacji). Atrybuty informacji mogą pozwolić mierzyć jakość informacji, jeżeli będą to miary ilościowe z uwzględnieniem odbiorcy tej informacji (Abramowicz, 2008). Warto zaznaczyć, że informacyjne atrybuty często noszą charakter subiektywny.

W literaturze istnieją różne podejścia do definiowania atrybutów informacji. Na przykład M. J. Eppler (2006) zaproponował 70 atrybutów informacji, które zawęża do 16 najistotniejszych. Przy ocenianiu subiektywnym, szczególnie ważne jest precyzyjne definiowanie tych atrybutów, żeby oceniający jasno rozumiał, co podlega ocenie (Grudzień, 2012).

Wiarygodność informacji można rozpatrywać w kontekście jej pochodzenia (informatora) oraz jako atrybut pochodny innych wartości (Swoboda, 2015). Jeżeli chcemy rozpatrywać wiarygodność w kontekście pochodzenia informacji, trzeba dokonać ocenę nadawcy informacji (źródła). Najczęściej zwraca się uwagę na kompetencje i kwalifikacje nadawcy w określonym obszarze wiedzy lub działalności. Wiarygodny nadawca ma większy stopień zaufania, jeżeli przekazał dużą ilość informacji z określonej dziedziny, przy popełnieniu minimalnej liczby błędów (Boruszewski, 2012).

Aktualność jest jednym z ważnych atrybutów informacji w gospodarce rynkowej: nieaktualne dane mogą spowodować podjęcie mniej efektywnych decyzji, straty finansowe, zmniejszenie zadowolenia klienta oraz utratę potencjalnych klientów (Experian QAS, 2013).

W ocenie jakości informacji umieszczonej w Internecie może pomóc tzw. metoda stosowania automatycznych procedur (Swoboda, 2015). W tym przypadku mogą być wykorzystane narzędzia do automatycznej weryfikacji technicznych atrybutów dokumentu. Błędy w kodzie mogą wskazać na gorszą jakość informacji w nim zawartych. Grijzenhout i Marx w swojej pracy stosując tylko zapytania XML wraz z językami transformacji wykryli prawie 15% błędnych dokumentów (Grijzenhout i Marx, 2013).

Określenie wymiarów jakości informacji zależy od miejsca, w którym ta informacja została umieszczona, np. informacja umieszczona w drukowanej książce i informacja umieszczonej na

portalu internetowym. Dla takiego źródła jak encyklopedia, w jego tradycyjnym przedstawieniu, zostało zdefiniowanych 7 wymiarów jakości (Crawford, 2001): format, unikatowość, autorytet, zakres, dokładność, obiektywność, aktualność. Dodatkowo, zestaw wymiarów jakości informacji może się różnić w zależności od typu ocenianej informacji - np. mapy geograficzne, obrazki, teksty oraz inne (Batini i Scannapieco, 2016).

Dodatkowym problemem jest brak ogólnie przyjętego zestawu wymiarów jakości - ich liczebność może być nawet większa niż 40 (M. Eppler, 2013; Schaal, Smyth, Mueller i MacLean, 2012). Jednocześnie należy brać pod uwagę, że takie obszerne listy wymiarów mogą i muszą być znacząca skrócone, ponieważ zawierają one różne niespójności i powtórzenia (M. Eppler, 2013). Niektóre wymiary mogą być wynikiem innych: np. wiarygodność, autorytet lub reputacja mogą być wynikiem analizy poprawności i spójności informacji. Dodatkowo te wymiary muszą być przydatne do użycia w praktyce. Dlatego, niektórzy badacze posługują się tylko 6-8 wymiarami przy ocenie jakości danych oraz informacji (Batini i Scannapieco, 2016; Crawford, 2001; S. Kim i Stoel, 2004).

2.4 Podsumowanie

Istnieją różnice pomiędzy definiowaniem jakości danych oraz jakości informacji. Jakość składa się z wymiarów, które są definiowane w zależności od kontekstu oraz źródła informacji. Przykładem takich wymiarów mogą być m.in. aktualność, wiarygodność, styl, kompletność oraz inne. Wymiary mogą mieć różne nazwy oraz zawierać określony zestaw miar. Wcześniejsze badania pokazały, że przy definiowaniu dużej liczby wymiarów (ponad 40) należy rozważyć ich redukcję, ponieważ niektóre z nich mogą być ze sobą powiązane.

W następnym rozdziale zostaną przedstawione serwisy wiki oraz najbardziej znany ich przykład – wolna encyklopedia Wikipedia. Zrozumienie kontekstu działania takich serwisów pomoże zdefiniować zestaw miar oraz wymiarów jakości odpowiednich do tego źródła informacji.

Rozdział 3

Serwisy wiki oraz semantyczne bazy wiedzy

W niniejszym rozdziale zostały opisane zagadnienia związane z serwisami wiki oraz jednym z popularnych przykładów tych serwisów - Wikipedią. Dodatkowo zostały opisane infoboksy, które zazwyczaj są umieszczane w widocznej części artykułu tej encyklopedii oraz które są wykorzystywane do prezentacji najważniejszych faktów. Opisana została również semantyczna baza DBpedia, która automatycznie wzbogacana danymi z tych infoboksów.

Obszerne fragmenty niniejszego rozdziału zostały opublikowane w pracach (Lewoniewski, Kasprzak, Węcel i Abramowicz, 2018; Lewoniewski, Węcel i Abramowicz, 2017c)

3.1 Wprowadzenie

Dopiero ok. roku 2001 zaczęły powstawać narzędzia i serwisy, które pozwoliły zwykłym użytkownikom przeglądarek na dostarczanie treści – pojawił się trend Web 2.0. Zmienił on sposób interakcji między właścicielami serwisu i jego użytkownikami, oddając tworzenie większości treści w ręce tych drugich. Społeczny charakter usług Web 2.0 oferuje praktycznie wszystkim użytkownikom możliwość swobodnego współtworzenia treści.

Nowe aplikacje Web 2.0 zapewniają nowe możliwości i kanały współdzielenia wiedzy. Na przykład blogi internetowe tworzone przez tysiące osób, serwisy wiki, które umożliwiają publiczne wspólne tworzenie informacji, technologii RSS, które za pomocą XML klasyfikują i organizują informacje (Deng i Luo, 2007). Te i inne możliwości sprawiają, że informacje w sieci są bardziej dostępne.

Otwarta treść (z ang. „open content”) jest to dowolny rodzaj wyniku pracy twórczej publikowany i licencjonowany w sposób umożliwiający swobodne wykorzystywanie przedmiotu

licencji. To pojęcie wraz z technologiami Web 2.0 pozwoliły na powstanie zjawiska crowdsourcingu, który został opisany dalej.

Model tworzenia treści przez rozproszoną społeczność szybko zmaterializował się jako model biznesowy. Firmy mogły zlecać internautom realizację określonego zadania, które było zbyt duże dla pojedynczego zespołu czy nawet firmy. Poprzez analogię do outsourcingu zjawisko to zostało określone jako crowdsourcing. Wśród projektów crowdsourcingowych wiele jest takich, które nie tylko stały się podstawą działania firm, ale również stały się rozpoznawalne globalnie. Na przykład Kaggle to platforma do organizacji konkursów z analityki danych, Khan Academy to organizacja non profit, której misją jest dostarczanie za darmo materiałów edukacyjnych najwyższej jakości, OpenStreetMap to darmowa mapa świata tworzona przez internautów, Waze to aplikacja mobilna wykorzystująca crowdsourcing do zbierania informacji o ruchu i zdarzeniach drogowych.

Najbardziej znanym przykładem źródła współtworzonego przez wiele osób jest Wikipedia. Zgodnie z jej zasadami informacja może być dostarczana przez każdego, również przez anonimowych użytkowników. Wikipedia jest popularnym przykładem serwisów wiki i często uważa się ją za projekt crowdsourcingowy (Buecheler, Sieg, Füchslin i Pfeifer, 2010). Wikipedia - szczególny przypadek serwisów wiki, które zostaną opisane w następnym podroździale.

3.2 Serwisy wiki

O serwisach wiki mówimy w przypadku zbioru powiązanych ze sobą stron internetowych, w których treść można tworzyć i zmieniać z poziomu przeglądarki internetowej. Strony wiki mogą być wykorzystywane do pracy nad wspólnymi projektami.

Strony typu wiki to również platforma wymiany wiedzy. Ta platforma promuje kreację wiedzy poprzez wzajemną współpracę (Jiao i Yuan, 2008). Różnica między wiki a blogiem polega na tym, że witryny wiki są zaprojektowane do współpracy między grupami użytkowników. Każdy może w każdej chwili edytować treść na wiki. W ramach takich serwisów dostępne również fora dyskusyjne dla każdej strony, umożliwiając użytkownikom prowadzenie rozmów na temat ich stron.

Jedną z głównych organizacji, która zarządza różnymi serwisami typu wiki jest Fundacja Wikimedia. Obecnie prowadzi ona 702 aktywne projekty w różnych językach (Wikipedia Meta-

Tabela 3.1. Lista 20 największych projektów fundacji Wikimedia pod kątem liczby artykułów z uwzględnieniem wersji językowych.

Projekt	Język	Artykuły	Strony	Edycje	Uzytkownicy
commons.wikimedia		48 878 991	66 380 600	317 353 404	7 205 568
en.wiktionary	Angielski	5 748 814	6 359 655	50 219 716	3 456 092
en.wikipedia	Angielski	5 709 436	45 759 359	852 931 885	34 376 843
ceb.wikipedia	Cebuański	5 381 698	8 964 659	23 687 176	50 728
mg.wiktionary	Malgaski	5 099 433	5 185 980	25 938 861	6 793
sv.wikipedia	Szwedzki	3 771 367	7 690 453	43 411 792	628 559
fr.wiktionary	Francuski	3 334 283	3 630 632	25 459 986	234 489
de.wikipedia	Niemiecki	2 215 479	6 211 862	179 396 475	2 987 641
fr.wikipedia	Francuski	2 037 063	9 695 952	151 378 202	3 197 547
nl.wikipedia	Holenderski	1 940 752	3 994 645	52 019 551	939 362
ru.wikipedia	Rosyjski	1 494 050	5 775 523	94 537 829	2 359 692
es.wikipedia	Hiszpański	1 467 398	6 451 679	109 908 672	5 122 582
it.wikipedia	Włoski	1 458 461	5 908 572	99 128 529	1 707 762
pl.wikipedia	Polski	1 297 520	2 895 965	54 191 992	906 440
war.wikipedia	Warajski	1 263 158	2 876 525	6 193 052	37 830
vi.wikipedia	Wietnamski	1 187 698	13 617 710	42 228 218	618 196
ja.wikipedia	Japoński	1 118 856	3 315 257	69 582 586	1 373 266
zh.wikipedia	Chiński	1 020 594	5 507 109	50 820 337	2 574 037
pt.wikipedia	Portugalski	1 004 205	4 693 072	52 855 502	2 140 446
ru.wiktionary	Rosyjski	982 771	1 416 751	9 970 031	202 151

Źródło: (Wikipedia Meta-Wiki, 2018a)

Wiki, 2018a). Tabela 3.2 przedstawia 20 największych projektów fundacji Wikipedia z uwzględnieniem wersji językowych.

Serwisy wiki mogą działać nie tylko jako ogólnodostępne zbiory informacji, ale również jako korporacyjne bazy wiedzy. Takie serwisy mogą zapewniać większą przejrzystość wszystkim procesom w organizacjach czy firmach. Dodatkowo serwisy wiki umożliwiają pracownikom współpracę w zakresie komunikacji oraz dzielenia się informacją, aktywnie proponować oraz rozwijać nowe idei. Poza tym, wiki umożliwia łączenie pracowników firmy oraz szybko zidentyfikować ekspertów z różnych dziedzin.

Istnieje wiele możliwości tworzenie własnych serwisów wiki przy pomocy ogólnodostępnego oprogramowania. Niżej lista z opisem niektórych z nich:

- **MediaWiki**[1] - to platforma zbudowana przy użyciu języka PHP dla dużych projektów. W rzeczywistości jest to oprogramowanie, na którym działa Wikipedia i inne projekty Fundacii Wikimedia[2].
- **Tiki**[3] - platforma do utworzenia serwisów wiki, która została pobrana ponad milion razy przez firmy, rządy, organizacje non-profit i osoby na całym świecie. Popularność plat-

[1] https://www.mediawiki.org
[2] https://wikimediafoundation.org/our-work/wikimedia-projects/
[3] https://tiki.org

Tabela 3.3. Lista 20 największych projektów w ramach serwisu Wikia pod kątem liczby artykułów.

Projekt	Artykuły	Strony	Edycje	Uzytkownicy
respuestas.wikia	2 483 083	4 486 350	5 720 974	15 441 276
colors.wikia	2 242 280	2 242 981	2 245 234	16 580 738
lyrics.wikia	2 010 752	3 224 161	31 599 990	15 530 860
answers.wikia	1 143 393	2 079 160	7 123 589	15 430 860
speedydeletion.wikia	766 115	1 050 302	1 148 994	15 430 860
lt.biologija.wikia	583 411	1 392 322	2 075 290	16 374 841
techteam-qa6.wikia	344 179	589 526	1 032 498	7 903 436
scratchpad.wikia	282 887	449 211	2 521 285	15 420 104
familypedia.wikia	246 415	619 025	1 403 716	15 530 860
military.wikia	240 593	652 830	4 160 249	15 441 276
frag.wikia	227 211	494 517	839 458	15 430 860
marvel.wikia	209 329	1 036 637	4 588 626	15 441 276
ru.vlab.wikia	204 717	314 798	476 486	16 626 422
respostas.wikia	192 653	512 139	836 557	16 638 441
eq2.wikia	178 678	287 332	897 389	15 420 104
reponses.wikia	178 134	468 287	1 113 583	15 463 788
crossgencomicsdatabase.wikia	147 165	302 818	1 507 286	8 662 262
starwars.wikia	144 303	476 583	7 799 848	15 441 276
prowrestling.wikia	110 142	437 341	1 440 067	15 420 104
yugioh.wikia	107 584	557 680	3 969 504	15 441 276

Źródło: (WikiStats, 2018)

formy wynika głównie z tego, że jest to coś więcej niż platforma wiki - umożliwia także tworzenie stron blogów, forów, kanałów RSS oraz ankiet.

- **DokuWiki**[4] - oprogramowanie wiki, którego możliwości są bardzo zbliżone do Tiki i MediaWiki, pomimo braku niektórych zaawansowanych funkcji. Najważnieszą zaletą DokuWiki jest łatwość użytkowania.

Do utworzenia własnych serwisów wiki można skorzystać ze specjalnych serwisów, które udostępniają platformę oraz serwer do działania tej platformy za darmo. Jednym z takich serwisów jest Wikia[5]. Ten serwis obecnie znajduje się w rankingu 50 najczęściej odwiedzanych stron internetowych na świecie (Alexa, 2018). Za pośrednictwem Wikia działa ponad 385 tys. różnych encyklopedii, w których ogólna liczba stron wynosi ponad 50 mln (Fandom, 2018). Najczęściej w ramach danego serwisu tworzone są bazy wiedzy na temat gier wideo, filmów, muzyki, komiksów. W tabeli 3.4 pokazana lista 20 największych projektów w ramach serwisu Wikia pod kątem liczby artykułów.

Innym przykładem serwisu, który umożliwia stworzenie własnych encyklopedii jest Gamepedia[6]. Serwis umożliwia tworzenie baz wiedzy na temat gier wideo. Obecnie zawiera ponad

[4] https://www.dokuwiki.org
[5] http://wikia.com
[6] https://www.gamepedia.com/

2000 różnych encyklopedii, z ponad 5 mln artykułami edytowanymi przez ponad 1,2 mln użytkowników.

Niektóre serwisy wiki działają na własnych platformach. Na przykład Baidu Bake[7] - chińska encyklopedia posiadająca ponad 15 mln stron, które były edytowane ponad 144 mln razy. Całkowita liczba zarejestrowanych użytkowników w tym serwisie to ponad 6,5 mln.

3.3 Wikipedia jako przykład serwisu wiki

Przez 15 lat od czasu powstania Wikipedia zdobyła pozycję jednego z ważniejszych źródeł ogólnodostępnej informacji encyklopedycznej. Jej cechą charakterystyczną jest to, że jest ona współtworzona przez wielu użytkowników. Obecnie Wikipedia jest na piątym miejscu w rankingu najczęściej odwiedzanych stron w Internecie, ustępując tylko Google, YouTube, Facebook oraz Baidu.

Koncepcja Wikipedii jest dość prosta: otwarta encyklopedia, którą może edytować każdy. Została ona uruchomiona 15 stycznia 2001 roku. Wikipedia jest stworzona przede wszystkim dla ludzi, które chcą lepiej poznać soją historię, społeczeństwo oraz kulturę. Zarówno instytucje badawcze jak i firmy mogą nieodpłatnie korzystać z tej encyklopedii do poszerzenia wiedzy oraz polepszenia technologii.

Obecnie Wikipedia zawiera ponad 48 mln artykułów w około 300 różnych językach[8]. Największa jest angielska (EN) wersja językowa, która zawiera ponad 5,5 mln artykułów. Do jednych z najbardziej rozwiniętych wersji językowych należą również niemiecka (DE) z ponad 2 mln artykułami, a także francuska (FR), rosyjska (RU), polska (PL) z ponad 1 mln artykułów każda. W Wikipedii obok informacji na temat znanych osób, miast czy wydarzeń można również znaleźć treści związane z produktami takimi, jak filmy, samochody, telefony komórkowe. Każdy produkt może być opisany w różnych językach.

Zmiany wprowadzane przez użytkowników do każdego artykułu są zapisywane w historii edycji, która pozwala na śledzenie zmian i umożliwia przywrócenie zawartości artykułu do poprzedniej wersji. Do sierpnia 2018 użytkownicy dokonali łącznie ponad 2,4 miliarda edycji we wszystkich wersjach językowych.[9]

[7] https://baike.baidu.com/
[8] https://meta.wikimedia.org/wiki/List_of_Wikipedias
[9] https://meta.wikimedia.org/wiki/List_of_Wikipedias

W odróżnieniu od innych popularnych serwisów internetowych Wikipedia nie wyświetla reklam i utrzymuje się z darowizn od użytkowników. Według niektórych szacunków Wikipedia mogłaby zarobić na reklamie ponad 2300 mln USD rocznie[10]. W 2016 roku koszty utrzymania tej encyklopedii wyniosły około 66 mln USD, podczas gdy przychód z darowizn wyniósł ponad 77 mln USD[11].

Pomimo niekomercyjnego charakteru Wikipedii informacje zawarte w tej bazie wiedzy tworzonej przez społeczność mogą wpływać na decyzje biznesowe i konsumenckie. Strony Wikipedii o znanych osobach, firmach, produktach często pojawiają się jako pierwsze w wynikach wyszukiwania Google, Bing, Yandex i innych popularnych serwisów.

Artykuły o określonych produktach mogą powstawać niezależnie w każdej wersji językowej Wikipedii. W związku z tym jakość informacji o tym samym produkcie może się różnić w zależności od języka językami. Należy także zaznaczyć, iż opis produktu w jednej wersji językowej Wikipedii nie musi być zgodny z informacją zapisanej w innym języku.

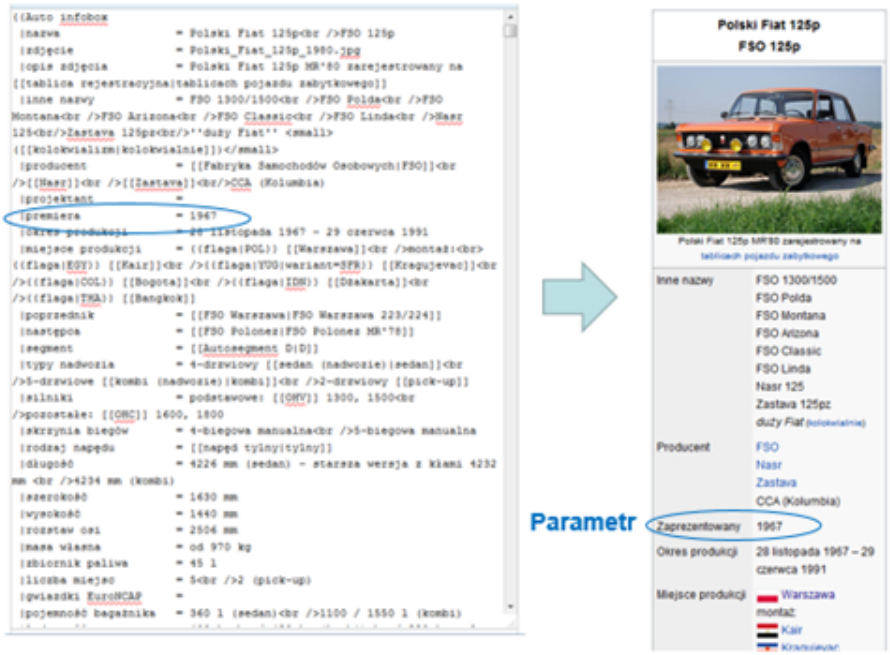
3.4 Infoboksy

Często w artykułach Wikipedii umieszcza się wyróżnioną ramkę, która w przejrzysty sposób ma prezentować najważniejsze informacje o podmiocie artykułu, tzw. infoboks. Czytelnicy korzystają z takich ramek, aby uzyskać najważniejsze fakty o produkcie bez analizy treści całego artykułu.

Od strony technicznej infoboks to szablon, który jest definiowany przez użytkowników Wikipedii, a następnie, odpowiednio wypełniony, umieszczany w artykule. To pozwala zachować spójny wygląd infoboksów danego typu w poszczególnych wersjach językowych Wikipedii. Zmiana w kodzie szablonu automatycznie pociąga za sobą zmiany we wszystkich artykułach z niego korzystających.

Szablon infoboksu składa się z elementów dwóch rodzajów: parametry i wartość parametrów. Infoboks określonego typu ma ściśle określony zestaw parametrów, które można używać do opisu określonego podmiotu czy wydarzenia. Na przykład w infoboksie opisującym samochody można użyć parametru „zbiornik paliwa”, natomiast w infoboksie o telefonach komórko-

[10] https://monetizepros.com/features/analysis-how-wikipedia-could-make-2-8-billion-in-annual-revenue
[11] https://upload.wikimedia.org/wikipedia/foundation/4/43/Wikimedia_Foundation_Audit_Report_-_FY15-16.pdf



Rysunek 3.1. Infoboks opisujący samochód (z lewej strony – kod źródłowy dla osoby edytującej artykuł, z prawej – wersja dla czytelników Wikipedii)

Źródło: Opracowanie własne na podstawie danych z Wikipedii.

wych ten parametr zostanie zignorowany i nie będzie wyświetlony czytelnikom Wikipedii. Na rys. 3.1 przedstawiony został przykład wypełnionego infoboksu o samochodzie.

Wstawianie infoboksu do artykułu nie jest obowiązkowe, ale jest zalecane dla stron określonego typu, dla których infoboksy zostały przewidziane.

Parametry infoboksów i ich wartości zazwyczaj wprowadzane są przez użytkowników, którzy mają różne doświadczenie i wiedzę na określony temat. Zatem wymagane jest dodatkowe sprawdzenie jakości tych danych przez bardziej doświadczonych redaktorów.

Niektóre wartości parametrów mogą pochodzić z innych źródeł niż kod artykułu, w którym jest umieszczony rozpatrywany infoboks. Jednym z takich źródeł jest projekt Wikidane[12] (z ang. Wikidata), który powstał w 2012 roku. Podobnie jak in Wikipedia, baza danych Wikidane jest projektem Wikimedia Foundation. Głównym zastosowaniem tego projektu jest używanie umieszczonych tam danych w Wikipedii oraz innych projektach – niektóre parametry określonych infoboksów mogą być uzupełniane automatycznie na podstawie tej bazy danych.

Kolejne źródło, z którego mogą być automatycznie wstawiane wartości do infoboksu to dane tabelaryczne (z ang. Tabular Data). Dane w tym przypadku przechowywane są na oddzielnej specjalnej stronie Wikipedii. W celu ekstrakcji danych z tej strony, zostają definiowane

[12] https://www.wikidata.org/

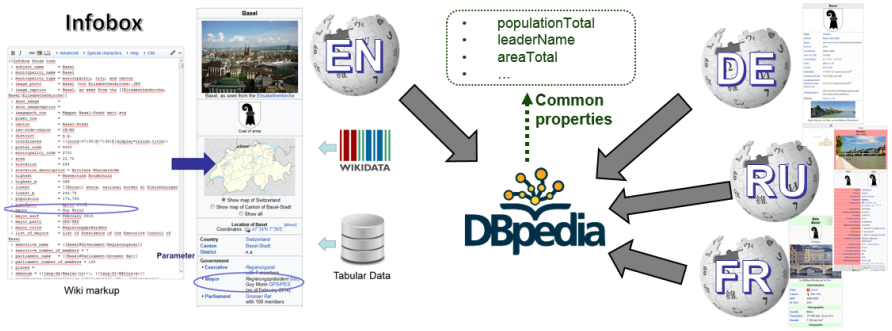
szablony, które wstawiane są potem do infoboksów w miejscu, gdzie należy wpisać wartość określonego parametru. Przy wykonaniu tego szablonu dane zostają ekstrahowane z tej specjalnej strony. Jest to szczególnie wygodne, gdy należy przeprowadzić aktualizacje podobnych typów parametrów (np. liczba ludności) dla infoboksów z artykułów o podobnej tematyce (np. miasta określonego państwa).

3.5 DBpedia

Infoboksy określonego typu często mają swoje odpowiedniki w różnych językach. Na przykład infoboks opisujący samochody w polskiej wersji ma nazwę „Auto infobox". Jego odpowiednik w niemieckiej to „Infobox PKW-Modell", a w angielskiej – „Infobox automobile". Różnice również można zauważyć w zestawie oraz nazwach parametrów. To może utrudniać porównanie wartości parametrów infoboksu pomiędzy różnymi wersjami językowymi tego samego produktu.

Pomocna w rozwiązaniu tego problemu może być semantyczna baza wiedzy – DBpedia, która odwzorowuje parametry infoboksów na specjalną ontologię i tym samym umożliwia wskazywanie ekwiwalentnych parametrów w różnych językach (Bizer i in., 2009). Warunkiem wstępnym jest poprawny opis reguł mapowania każdej nazwy parametrów w każdej wersji językowej. Na przykład dla samochodów parametry „producent" w polskiej wersji, „Marke" w niemieckiej, „manufacturer" w angielskiej będą mapowane do wspólnego „manufacturer" w ontologii DBpedii. Ta ontologia posiada ponad 300 klas z ponad 1600 różnymi opisanymi właściwościami (Lehmann i in., 2015).

Jakość infoboksów Wikipedii jest znacznie mniej rozwiniętym tematem w badaniach niż jakość artykułów. Istniejące badania często badają jakość baz danych tworzonych na podstawie ekstrahowanych informacji z infoboksów. Dobrym przykładem takich baz danych jest DBpedia, która dodatkowo zawiera wiele linków do innych zestawów danych w chmurze LOD (Linked Open Data), takich jak Freebase, OpenCyc oraz inne (Färber i in., 2016). Korzystając z kompleksowego zestawu ogólnych testów wzorcowych, można ujawnić znaczną ilość problemów z jakością danych (Kontokostas i in., 2014). Za pomocą specjalnych metod można analizować spójność, prawidłowość syntaktyczną, dokładność semantyczną i zwięzłość danych zawartych w DBpedii (Mihindukulasooriya i in., 2015). Analiza jakości danych w tej semantycznej bazie wiedzy jest również możliwa bez użycia ontologii (Jang i in., 2015). Istnieją również badania związane z połączeniem danych z różnych wersji językowych DBpedia (Tacchini i in., 2009). Jed-



Rysunek 3.2. Infobox o mieście Bazylea z jej źródłami danych oraz ekstrakcji danych do DBpedii z różnych wersji językowych Wikipedii.

Źródło: Opracowanie własne.

nak większość prac nie bierze pod uwagę różnych aspektów jakości infoboxów oraz stron wiki, z których te dane pochodzą.

Przy pomocy ontologii i systemu mapowań[13] DBpedia umożliwia unifikację nazw parametrów infoboksów na podobny temat w różnych wersjach językowych. To pozwala na późniejsze porównywanie wartości parametrów w różnych wersjach językowych (patrz rys. 3.2)

Istnieją badania, które pokazują w sposób automatycznej aktualizacji danych w infoboksach na podstawie danych z różnych semantycznych baz danych, takich jak DBpedia (Ahmeti, Fernández, Polleres i Savenkov, 2017) czy Wikidata (Sáez i Hogan, 2018).

3.6 Podsumowanie

Rozwój technologii internetowych pozwala na coraz większe zaangażowanie użytkowników internetu we współtworzenia baz wiedzy. Systemy współtworzenia treści stają się coraz bardziej dostępne, co zwiększa liczbę potencjalnych autorów ogólnodostępnych treści. Jednym z popularnych narzędzi współtworzenia wiedzy są serwisy wiki. Używając rożnych platform można stworzyć kompleksową bazę wiedzy na potrzeby własne, korporacyjne lub społeczne.

Popularnym przykładem serwisów wiki jest Wikipedia, która została opisana w niniejszym rozdziale. Dodatkowo zostały scharakteryzowane infoboksy, które pozwalają na wygodne przedstawienie danych o podmiocie w artykułach Wikipedii oraz są szczególne istotne w ramach tej rozprawy. Infoboksy mogą być również wykorzystane do tworzenia semantycznych baz wiedzy, takich jak DBpedia.

[13] `http://mappings.dbpedia.org`

W następnym rozdziale zostaną poruszone kwestie jakości w tej encyklopedii oraz zostanie opisane, w jaki sposób można zautomatyzować proces oceny jakości artykułów.

Rozdział 4

Metody określenia jakości artykułów Wikipedii

W niniejszym rozdziale zostały przedstawione zagadnienia związane z definiowaniem wymiarów jakości w zależności od źródła ocenianej informacji. W szczególności zostaną zdefiniowane wymiary jakości artykułów Wikipedii. Rozdział opisuje również problemy związane z jakością treści spotykane w Wikipedii. Dodatkowo zostały opisane dostępne metody automatycznego określenia jakości artykułów tej encyklopedii.

4.1 Wprowadzenie

Współtworzenie informacji w Wikipedii może się wiązać z różnymi problemami, które wpływają na jakość treści.

Kluczową kwestię stanowi jakość treści, która często jest zależna od tematu i wersji językowej Wikipedii. Istnieje szereg badań pomagających w automatycznej estymacji jakości artykułów. W tym rozdziale zostały opisane znane już podejścia do automatycznej oceny jakości artykułów przy pomocy różnorodnych modeli wykorzystujących cechy artykułów. Jakość artykułów serwisów można określić poprzez ocenę m.in. następujących wymiarów: aktualność, kompletność, wiarygodność. Wymiar jakości to odpowiednio dobrany zbiór miar. Miara to wartość ilościowa, wyliczona na podstawie określonych reguł. Na przykład kompletność można określić na podstawie długości tekstu lub liczby nagłówków.

4.2 Wymiary jakości serwisów wiki

W celu zdefiniowania wymiarów jakości serwisów wiki należy brać pod uwagę podobieństwo tych serwisów z tradycyjnymi encyklopediami oraz dokumentów Web 2.0. Większość dotychczasowych badań w zakresie jakości informacji w serwisach wiki skupiają się wokół najbardziej popularnego przedstawiciela tychże serwisów - Wikipedii. Wyniki badań pokazują, że treść współtworzona przez użytkowników w Wikipedii może być uznana za encyklopedyczną, ponieważ ma taką samą dokładność, jak w przypadku tradycyjnych encyklopedii (Giles, 2005).

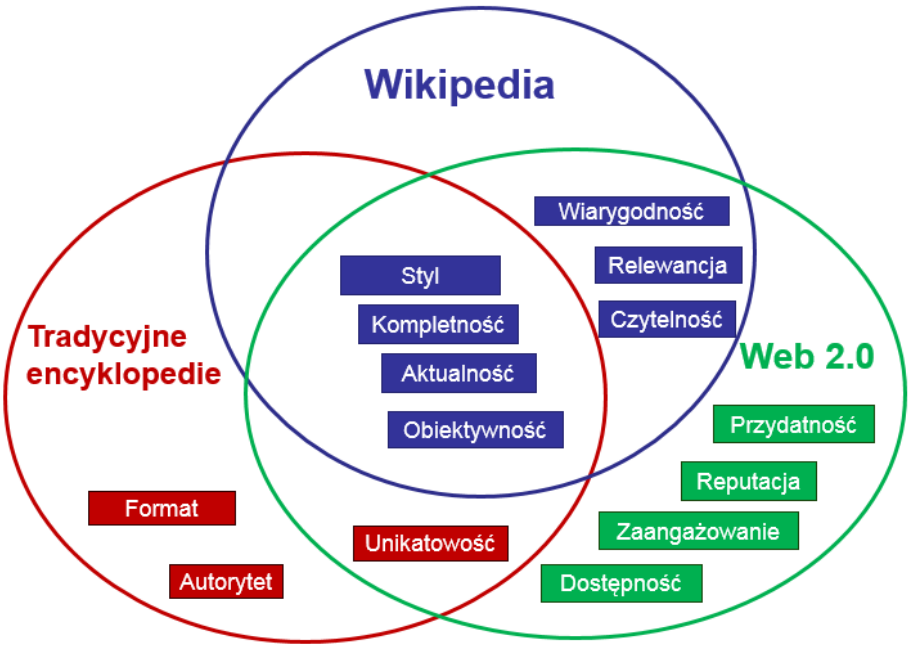
Jakość artykułu w tradycyjnej encyklopedii może być definiowana poprzez 7 ogólnych wymiarów (Crawford, 2001):

- Format – na jakim nośniku przedstawiona informacja.
- Unikatowość – zawieranie cech, które odróżniają ją od innych encyklopedii.
- Autorytet – reputacja osób, które sprawdzają informację (recenzenci).
- Zakres – jakie tematyczne obszary pokrywa, na jakiego użytkownika skierowana, w jakim stylu napisana.
- Dokładność – czy zawiera błędy.
- Obiektywność – wszechstronność i neutralność materiału, zawieranie obrazków i innych materiałów multimedialnych.
- Aktualność – zgodność z rzeczywistością w momencie, gdy jest użytkowana.

Powyższe kryteria częściowo pokrywają się z kryteriami ustalonymi przez społeczność Wikipedii. Autorytet w mniejszym stopniu dotyczy Wikipedii – artykuły nie muszą być sprawdzane przez ekspertów. Ważnymi w Wikipedii natomiast są takie elementy jak: neutralny punkt widzenia, powoływanie się na oryginalne badania zamiast ich prezentacji, weryfikowalność.

Wikipedia, jako przedstawiciel serwisów wiki, działa dodatkowo jako encyklopedia, która funkcjonuje nieco na innych zasadach niż tradycyjne (drukowane). Na przykład, artykuły Wikipedii mogą być tworzone oraz redagowane przez użytkowników (w tym anonimowych) w czasie rzeczywistym, wówczas zmiany są widoczny od razu dla czytelników tej encyklopedii.

Serwisy wiki działają na zasadach Web 2.0 (Deng i Luo, 2007). Najważniejszą cechą takich dokumentów jest to, że ich treści mogą być generowane przez użytkowników Internetu (Dalip, Gonçalves, Cristo i Calado, 2017). W konsekwencji pojawiły się nowe rodzaje repozytoriów wiedzy, do współtworzenia których każdy może się swobodnie przyczynić. Przykładami takich repozytoriów mogą być serwisy oparte na mechanizmie pytania - odpowiedzi (Q&A), cyfrowe



Rysunek 4.1. Wymiary jakości dla informacji z tradycyjnych encyklopedii, dokumentów Web 2.0., Wikipedii.

Źródło: Opracowanie własne.

repozytorium wideo, blogowanie, witryny z recenzjami, serwisy społecznościowe oraz inne. Na podstawie literatury (Dalip i in., 2017; Herrera-Viedma, Pasi, Lopez-Herrera i Porcel, 2006; Mohanty, Seth i Mukadam, 2007; Pun i Lochovsky, 2004; Schaal i in., 2012) oraz własnych obserwacji, dla oceny jakości dokumentów Web 2.0 zostały zdefiniowane następujące wymiary: unikatowość, styl, kompletność, aktualność, obiektywność, weryfikowalność, relewancja„ czytelność, przydatność, reputacja, zaangażowanie.

Pokrycie wymiarów jakości dla informacji z tradycyjnych encyklopedii, dokumentów Web 2.0 i Wikipedii pokazane jest na rys. 4.1. Dla każdego z wymiarów istnieje odpowiedni zbiór miar. Przykład jest pokazany w tabeli 4.1.

4.3 Problemy jakości Wikipedii

Wikipedia nie ma centralnej redakcji czy grupy użytkowników, którzy mogliby metodycznie oceniać artykuły. Byłoby to tym bardziej trudne, że treść jest dynamiczna – może zmieniać się nawet kilka razy dziennie. Co więcej, autorzy nie muszą formalnie potwierdzać swoich umiejętności w określonej dziedzinie. Te i inne aspekty funkcjonowania Wikipedii były przedmiotem krytyki, w szczególności wskazującej na słabą jakość informacji.

Tabela 4.1. Wymiary jakości serwisów wiki.

Wymiar	Opis	Sposób mierzenia
Aktualność	na ile artykuł opisuje obecny stan pewnej rzeczywistości	Data ostatniej modyfikacji artykułu, częstość wprowadzania zmian
Kompletność	na ile wyczerpujący jest opis tematu	Długość artykułu i poszczególnych jego elementów (abstraktu, tabeli, sekcji…).
Weryfikowalność	na ile przedstawione informacje można sprawdzić	Liczba referencji, jakość referencji, wskaźnik ref/długość, liczba referencji na sekcję, czy zawiera specjalne szablony wskazujące na problemy z weryfikowalnością
Styl	w jaki sposób jest organizowana zawartość artykułu	Liczba i kolejność sekcji, liczba tabel, liczba linków, liczba szablonów, liczba obrazków, czy zawiera specjalne szablony wskazujące na luki w stylu
Objektywność	na ile zawartość artykułu spełnia kryterium neutralnego punktu widzenia, czy zawiera obrazki i inne materiały multimedialne dotyczące tego artykułu	liczba obrazków i innych elementów multimedialnych dotyczące tego artykułu (np. nie liczone są obrazki ze znaczkami flag), specjalne szablony, analiza NLP
Czytelność	na ile tekst jest zrozumiały i wolny od niepotrzebnej złożoności	Wskaźnik Flesch-Kincaid, wskaźniki użycia różnych słów (w tym fachowych).
Relewancja	na ile artykuł jest ważny dla użytkowników	Liczba wyświetleń, liczba obserwujących użytkowników, ocena ważności przez Wiki-projekty, linki na artykuł, PageRank, liczba autorów

Źródło: Opracowanie własne.

Można się spodziewać, że osoby odwiedzające strony Wikipedii oraz jej twórcy są zainteresowani wysoką jakością treści w niej zawartej. Dotyczy to również dużych korporacji, które m.in. muszą dbać o obiektywne przestawienie informacji o swoich produktach. Na przykład w 2017 roku sieć restauracji Burger King przez swoją reklamę proponowała wyszukać w wyszukiwarce Google informacji na temat ich nowego produktu, który był również opisany w Wikipedii. W tym czasie opis tego produktu w Wikipedii został zmieniony przez jednego z użytkowników – we wprowadzeniu było zaznaczone, że jest to „najgorszy produkt", a także dodane inne niesprawdzone informacje (Wakefield, 2017).

Według Wikipedii jakość jej artykułu jest rozumiana jako zdolność artykułu do sprostania oczekiwaniom i potrzebom docelowych odbiorców tego artykułu (Wikimedia Strategic Planning, nodate).

Pojęcie jakości w Wikipedia nie jest stałe ze względu na czas i przestrzeń. Obecnie Wikipedia zawiera około 300 wersji językowych i każda ma swoją społeczność użytkowników, która może samodzielnie (niezależnie od innych języków) definiować pojęcie jakości w ramach swojej wersji (Lewoniewski, Węcel i Abramowicz, 2017b; Stvilia, Al-Faraj i Yi, 2009; Warncke-wang i in., 2013; Węcel i Lewoniewski, 2015). Systemy oceniania mają swoje ustalone w toku dyskusji i wieloletniej praktyki regulaminy i zasady przyznawania wyróżnień za jakość.

W Wikipedii istnieje system wyróżnień dla artykułów, które uważane są za wzorowe. Takie artykuły istnieją praktycznie w każdej wersji językowej. W angielskiej wersji Wikipedii najlepsze artykuły nazywają się Featured Article (FA). Istnieje również drugie wyróżnienie dla artykułów, które jeszcze nie spełniają wszystkich kryteriów FA, ale zbliżają się do ich jakości - Good Article (GA). W polskiej wersji odpowiednikami FA i GA są „Artykuł na medal" i „Dobry artykuł". Nadawanie tych wyróżnień następuje w wyniku dyskusji w drodze konsensusu, w którym każdy użytkownik może się zgodzić lub nie z nadaniem artykułowi wyższej oceny jakości oraz wyjaśnić swój punkt widzenia. Te zasady mogą się zmieniać w czasie, niezależnie w każdej wersji językowej, co w konsekwencji może spowodować utratę wyróżnienia przez niektóre artykuły.

W niektórych wersjach językowych artykuł może dostać również inną ocenę jakości. Taka ocena może pokazać, na ile „dojrzały" jest artykuł oraz w jakim stopniu jest on zbliżony do wzorowego. Na przykład w angielskiej wersji Wikipedii wyróżnia się 7 klas jakości artykułu (od najwyższej): FA, GA, A-class, B-class, C-class, Start, Stub. Wszystkie oceny poniżej FA i GA nadawane są bez dyskusji społeczności i uzyskania konsensusu – każdy użytkownik może wystawić ocenę samodzielnie na podstawie przyjętych zasad. W wersji polskiej oprócz wyróżnionych

istnieją jeszcze następujące oceny: Czwórka, Start, Zalążek. Warto dodać, że w zależności od projektów tematycznych w ramach tej samej wersji językowej mogą być stosowane rożne nazwy klas o podobnych kryteriach. Np. podobną do klasy „czwórka" w polskojęzycznej Wikipedii jest klasa o nazwie „poprawny", a odpowiednikiem klasy startowej w oddzielnych projektach tematycznych występuję klasa dostateczna.

Dodatkowym interesującym aspektem w ocenie jakości artykułów przez użytkowników jest to, że jeden artykuł może posiadać jednocześnie różne klasy jakości w zależności od projektu tematycznego. Na przykład na moment przeprowadzenia badań artykuł o Poznaniu w angielskiej Wikipedii[1] jest oceniony oddzielnie z punktu widzenia pięciu projektów tematycznych: WikiProject Poland, WikiProject Middle Ages, WikiProject Cities, WikiProject Former countries / Prussia oraz WikiProject Germany. W tym przypadku oceny pomiędzy projektami są spójne: klasa C. Oceny od poszczególnych projektów zazwyczaj można sprawdzić na stronach dyskusji nad artykułem. Przykładem niespójności ocen pomiędzy projektami tematycznymi można obserwować na przykładzie artykułu o Taiwanie w angielskiej Wikipedii[2], gdzie 5 projektów ocenia artykuł jako klasa B, podczas gdy inne 3 projekty przypisały klasę C. Artykuł może nawet mieć ponad 10 ocen jednocześnie od różnych projektów tematycznych - np. artykuł o Google w anglojęzycznej Wikipedii[3]. Dodatkowo każdy projekt może wyznaczyć ocenę ważności danego artykułu.

W tabeli 4.2 pokazana jest liczebność artykułów w poszczególnych klasach jakości. Dla artykułów, które miały niespójne oceny jakości pomiędzy projektami tematycznymi Wikipedii, była wybierana najniższa.

Z zestawienia wynika, że nie ma ogólnie przyjętego standardu klasyfikacji artykułów pomiędzy różnymi wersjami językowymi Wikipedii. Niektóre języki stosują rozwiniętą skalę ocen (EN, RU), inne ograniczają się do 2-3 klas jakości (BE, DE). Poza tym, w rozwiniętych klasyfikacjach pomiędzy językami również nie ma spójności, jednak można znaleźć podobieństwa w zasadach przyznawania poszczególnych ocen (w tabeli 4.2 podobne klasy zostały pogrupowane).

Każda wersja językowa może mieć swój system klasyfikacji jakości artykułów, jednak można zauważyć, że wszystkie stosują co najmniej dwie klasy wyróżnionych artykułów – odpowiedniki FA i GA. Takich artykułów jest bardzo mało – średnio w każdej wersji językowej udział ich wynosi około 0,07%. Warto również podkreślić, że duża część artykułów nie jest oceniona, np. w

[1]Adres strony dyskusji nad artykułem o Poznaniu w angielskiej Wikipedii: `https://en.wikipedia.org/wiki/Talk:Poznań`

[2]Adres strony dyskusji nad artykułem o Taiwaniu w angielskiej Wikipedii: `https://en.wikipedia.org/wiki/Talk:Taiwan`

[3]Adres strony dyskusji nad artykułem o Google w angielskiej Wikipedii: `https://en.wikipedia.org/wiki/Talk:Google`

Tabela 4.2. Liczba artykułów w poszczególnych klasach jakości w różnych wersjach językowych Wikipedii.

	Wersja językowa				
	BE	**EN**	**PL**	**RU**	**UK**
Wszystkie artykuły	155,256	5,674,716	1,288,046	1,481,270	798,850
Featured Article (FA)	66	5,344	795	1,119	247
Good Article (GA)	109	28,082	2,107	3,122	649
Solidny artykuł				4,977	
A-class		2,061			
Czwórka			167		
Pełny artykuł				5,585	191
B-class		73,081			
Rozwinięty artykuł				18,464	1,601
C-class		202,551			
Artykuł w rozwoju				76,901	7,017
Start		1,430,317	862		
Zalążek (Stub)	964	2,834,390	637	85,906	25,187
Nieocenione	154,117	1,098,890	1,283,478	1,285,196	763,958

Źródło: Opracowanie własne w lipcu 2018r.

polskiej edycji udział takich artykułów stanowi ponad 99% (Lewoniewski, Węcel i Abramowicz, 2017b).

4.4 Metody automatycznego określenia jakości artykułów Wikipedii

Od momentu powstania i w miarę wzrostu popularności Wikipedii pojawia się coraz więcej publikacji naukowych na temat jakości informacji w niej zamieszczanych.

Jedno z pierwszych badań pokazało, że pomiar objętości treści może pomóc w określeniu stopnia dojrzałości artykułu (Stvilia i in., 2005a). Prace w tym kierunku pokazują, że zazwyczaj artykuły wyższej jakości są dłuższe (Blumenstock, 2008b), wykorzystują w spójny sposób referencje, są edytowane przez setki redaktorów i posiadają tysiące edycji (Hu, Lim, Sun, Lauw i Vuong, 2007; Wöhner i Peters, 2009).

Oprócz analizy ilościowej późniejsze badania skupiały się również wokół analizy jakościowej treści artykułu. W jednej z prac został wykorzystany tzw. indeks czytelności FOG, który określa stopień przystępności tekstu (Dalip, Gonçalves, Cristo i Calado, 2009). Istnieją badania, które wykorzystują cechy lingwistyczne ekstrahowane z tekstu w celu analizy jakości artykułów. Lipka (Lipka i Stein, 2010) analizuje rozkład ciągów znaków (trigramów) w artykułach w celu automatycznej oceny jakości informacji. W innych badaniach zaproponowano wykorzystanie liczby faktów i gęstości faktów jako miar do identyfikacji artykułów wysokiej jakości w Wikipedii (Horn, Zhila, Gelbukh, Kern i Lex, 2013), przy czym fakt ma postać trójki „podmiot – orzeczenie – dopełnienie" (Lex i in., 2012).

Użytkownicy Wikipedii mogą wystawiać specjalne szablony do artykułu, wskazujące na luki w jakości. Takie adnotacje mogą pomóc w ocenie jakości artykułu (Anderka, 2013). Cechy dotyczące popularności artykułu mogą być również wykorzystane przy ocenie jakości informacji w nich zawartych (Lewoniewski i in., 2015).

Kolejne prace dotyczące automatycznej klasyfikacji jakości artykułów Wikipedii uwzględniają zachowania użytkowników. Istnieją modele, które biorą pod uwagę ich doświadczenie i reputację. Artykuły wysokiej jakości mają dużą liczbę edycji i dużą liczbę redaktorów, którzy charakteryzują się wysokim poziomem współpracy (Kittur i Kraut, 2008; Wilkinson i Huberman, 2007a). Ważne jest to, żeby w tej grupie redaktorów był chociażby jeden użytkownik z wysokim poziomem doświadczenia w edycji treści w Wikipedii (Arazy, 2010). Dodatkowo, jakość

artykułów edytowanych przez sprzeczających się użytkowników jest najczęściej o wiele niższa od jakości artykułów, których współautorzy starają się spojrzeć szerzej na problem (Jemielniak, 2013). Szczególne znaczenie ma reputacja użytkownika, który dokonał pierwszej edycji artykułu (Stein i Hess, 2007). Reputacja użytkownika może być liczona na podstawie „przetrwania" tekstu, który on umieścił (Adler i De Alfaro, 2007; Halfaker, Kraut i Riedl, 2009; Suzuki i Yoshikawa, 2012).

W niniejszej pracy budowanie modeli jakości będą się opierać bardziej na miarach dotyczących treści artykułu i jego metadanych niż analizie zachowania jego współtwórców. To może pomóc m.in. w odnalezieniu elementów, które należy dopracować w artykule. W tej rozprawie zostały wykorzystywane już znane z literatury miary artykułów i zaproponowane nowe, w celu zwiększenia precyzji modeli jakości.

Miary, które można brać pod uwagę przy określeniu jakości artykułów, są definiowane w literaturze w zależności od dostępnych technologii. Np. miary dotyczące popularności artykułu coraz częściej występują w pracach naukowych jako element składowy modeli estymacji jakości Wikipedii. Wcześniejsze badania wykazały, że w przypadku niektórych rozwiniętych wersji językowych Wikipedii (takich jak angielska, niemiecka i hiszpańska) popularność artykułów była skorelowana z liczbą edycji (Reinoso, 2011). Inne badania wykazały, że mierzenie popularności tematu w anglojęzycznej Wikipedii może pomóc w ustaleniu liczby artykułów dobrej jakości – jeśli temat jest popularny, to ma większą liczbę artykułów wysokiej jakości (Lehmann, Müller-Birn, Laniado, Lalmas i Kaltenbrunner, 2014). Z drugiej strony, Warncke-Wang i inni pokazali niedopasowanie między popularnością i jakością artykułów w Wikipedii; jednak badanie ograniczało się do czterech wersji językowych Wikipedii (Warncke-Wang, Ranjan, Terveen i Hecht, 2015). Ponadto, żadne z badań nie zawierało analizy porównawczej popularności tego samego artykułu między wersjami językowymi a jego wpływem na jakość. Popularność może również w pewnym stopniu pokazywać ważność artykułu w wybranej wersji językowej dla odpowiedniej grupy narodowej użytkowników Wikipedii.

Biorąc pod uwagę niejednoznaczność wyników wcześniejszych prac naukowych, w niniejszej pracy zostały przeprowadzone bardziej rozwinięte badania miar dotyczących popularności artykułów i ich wpływu na jakość informacji. To jest szczególne ważne z punktu widzenia porównania różnych wersji językowych tego samego artykułu — większa liczba użytkowników może sprawdzić kompletność, aktualność i wiarygodność faktów opisanych w artykule. Dodatkowo, w istniejących pracach popularność porównywana jest zazwyczaj z jakością całego artykułu, a

nie jakością poszczególnych jej części składowych (np. infoboksy). W niniejszej pracy również na ten aspekt zwrócono uwagę.

4.5 Podsumowanie

Badania naukowe dotyczące automatycznej oceny jakości artykułów Wikipedii są już stosunkowo zaawansowane, ciągle mają jednak swoje ograniczenia. Różne prace skupiają się zazwyczaj na miarach z określonych wymiarów jakości i wybranych wersjach językowych (najczęściej - angielskiej).

Nowe modele jakości wraz z włączeniem dodatkowych miar powstawały podczas rozwinięcia każdej wersji językowej Wikipedii wraz ze społecznością użytkowników, która ciągle wprowadza zmiany i udoskonala zasady oceny jakości artykułów. Należy też brać pod uwagę rozwój platformy MediaWiki, która zapewnia techniczne działanie tej popularnej encyklopedii. Nowe funkcje umożliwiają zdefiniowanie nowych miar i wymiarów jakości, o których jest mowa w następnych rozdziałach.

W związku z tym, można się spodziewać, że niektóre z opisanych we wcześniejszych pracach modeli jakości, miar jakości oraz metod porównywania jakości artykułów Wikipedii utraciły aktualność. Usprawnienie tych modeli oraz metod jest jednym z głównych celów niniejszej pracy.

Rozdział 5

Miary oraz wymiary jakości artykułów Wikipedii

W niniejszym rozdziale zostały przedstawione miary jakości artykułów Wikipedii, które zostały przypisane do odpowiednich wymiarów jakości. Te miary pozwoliły na zbudowanie modeli automatycznej ewaluacji jakości artykułów. Dobór miar jest wynikiem analizy podobnych prac w tym kierunku oraz własnych eksperymentów.

W celu odwołania się do poszczególnych miar jakości artykułów Wikipedii używane są skróty A_x (lub Ax), gdzie x - wartość liczbowa.

- A - oznacza miarę artykułu Wikipedii.
- x - wskazuje na indeks (numer) miary.

5.1 Wprowadzenie

Większość badań dotyczących budowania modeli jakości artykułów Wikipedii skupia się na największej wersji językowej – angielskiej. W niniejszej pracy zostało wzięte pod uwagę 55 najbardziej rozwiniętych wersji językowych. Pozwala to na budowanie modeli, które będą w stanie porównywać pomiędzy sobą jakość artykułów na jeden temat w różnych językach. Innymi słowy, celem niniejszego rozdziału jest zaproponowanie kompleksowego modelu jakości informacji, który miałby zastosowanie do względnej oceny jakości artykułów w Wikipedii.

Z jednej strony dane i informacje dostarczane przez społeczność mogą być kwestionowane – przykład wspomnianej już krytyki Wikipedii. Z drugiej strony stworzone ramy oceny jakości

przez społeczność pozwalają na poprawę jakości danych, w szczególności w tej części ustrukturyzowanej, tj. w infoboksach. Wykorzystane mogą być dwa mechanizmy:

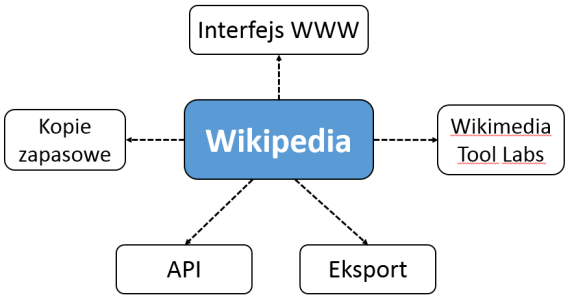
- formalny proces oceny jakości artykułów Wikipedii, który dostarcza nam danych,
- wielojęzyczność i powiązania między artykułami w różnych językach, stwarzające szanse w zakresie porównania i weryfikacji poprawności wprowadzanych danych.

5.2 Metody oraz źródła ekstrakcji miar

W związku z wielością źródeł, z których można pozyskać miary do oceny jakości informacji, zostaną zaproponowane metody adekwatne do każdego ze źródeł.

Najważniejsze metody pozyskania to: API, pobranie zrzutu stron (dump), zapytania SQL, zapytania SPARQL, przetwarzanie języka naturalnego (NLP). W przypadku NLP należy uwzględnić różne frameworki oraz wielojęzyczność, tzn. pozyskiwanie danych z wykorzystaniem charakterystycznych zasobów, np. do analizy morfologicznej.

Na rysunku 5.1 pokazane są różne podejścia do pozyskiwania danych dotyczących artykułów Wikipedii. Najprostszym jest interfejs WWW. Kolejnym stosunkowo prostym narzędziem dostępu do artykułów Wikipedii jest specjalny formularz[1] pozwalający eksportować jeden lub wiele artykułów w formacie XML wraz z historią edycji (do 1000 edycji).



Rysunek 5.1. Możliwości dostępu do danych artykułów Wikipedii

Źródło: Opracowanie własne.

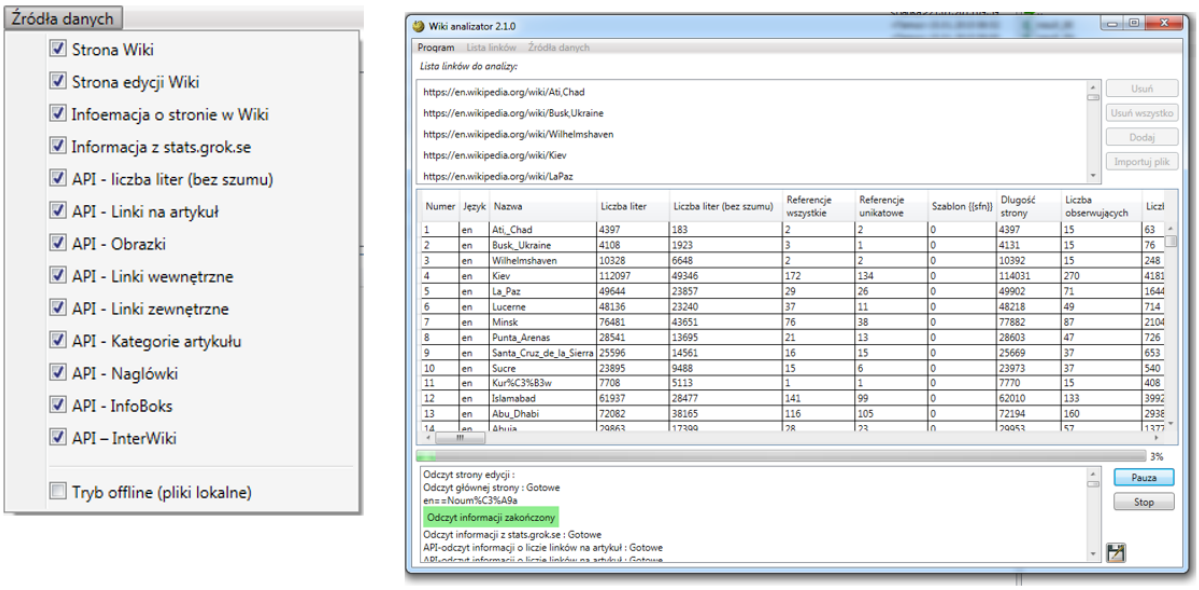
Wikipedia co miesiąc tworzy kompletną kopie wszystkich swoich wersji językowych w postaci wikitekstu (kodu źródłowego) i metadanych w różnych formatach (w tym XML) oraz surowych baz danych w postaci SQL.

[1] https://pl.wikipedia.org/wiki/Specjalna:Eksport

Wszystkie te pliki są swobodnie dostępne na specjalnym serwerze[2]. Wykorzystanie kopii zapasowych do analizy dużej liczby artykułów wygląda interesująco, jednak, w odróżnieniu od innych metod, jest bardziej czasochłonne. Jest to związane m.in z wielkością plików, np. dla angielskiej wersji Wikipedii jeden z plików zawierający ostatnie wersje tekstów źródłowych artykułów ma objętość ponad 70GB.

Wikipedia Tool Labs[3] (dawniej Toolserver) jest środowiskiem hostingowym dla użytkowników Wikipedii dla tworzenia różnego oprogramowania, które skierowane jest na ułatwienia korzystania z tej encyklopedii. Ponad 1000 narzędzi pozwala m.in. otrzymać różnego rodzaju statystyki dotyczące artykułów. Jednak programy te nie zawsze są dostępne, mogą zawierać błędy lub nieaktualne sposoby analizy informacji.

Jedną z najbardziej atrakcyjnych metod pozyskiwania danych jest serwis API, który zapewnia wygodny dostęp m.in. do danych i metadanych artykułów Wikipedii za pomocą protokołu HTTP, za pośrednictwem adresu URL, w różnych formatach (w tym XML, JSON). W odróżnieniu od kopii zapasowych, pobierane dane są aktualne na moment zapytania i odpowiedź serwera na zapytanie jest szybka. Z możliwości API korzysta specjalnie przygotowany dla naszych badań program WikiAnalizator, który może pozyskać ponad 50 różnych cech poszczególnych artykułów. Na rysunku 5.2 pokazany jest interfejs graficzny tego programu oraz źródła, z których pobiera on dane.



Rysunek 5.2. Interfejs graficzny programu WikiAnalizator wraz z wykazem źródeł danych.

Źródło: Opracowanie własne.

[2] https://dumps.wikimedia.org
[3] http://tools.wmflabs.org

Serwis API działa dla każdego języka i dostępny jest pod adresem określonym wg szablonu:

- `https://{język}.wikipedia.org/w/api.php?action={ustawienia}`, gdzie `{język}` – skrót wersji językowej,
- `{ustawienia}` – ustawienia zapytania[4].

Np. w celu otrzymania określonych informacji dotyczących artykułu „Polska" w polskim języku w formacie JSON należy wykorzystać następujące wywołania:

- kod źródłowy artykułu w notacji wiki:

  `https://pl.wikipedia.org/w/api.php?action=parse&format=json&prop=wikitext&page=Polska`

- tylko tekst artykułu:

  `https://pl.wikipedia.org/w/api.php?format=json&action=query&prop=extracts&explaintext&titles=Polska`

- lista wszystkich nagłówków:

  `https://pl.wikipedia.org/w/api.php?action=parse&format=json&prop=sections&page=Polska`

- lista wszystkich linków wewnętrznych:

  `https://pl.wikipedia.org/w/api.php?action=parse&format=json&prop=links&page=Polska`

- lista wszystkich szablonów:

  `https:///pl.wikipedia.org/w/api.php?action=parse&format=json&prop=templates&page=Polska`

- lista wszystkich obrazków:

  `https://pl.wikipedia.org/w/api.php?action=parse&format=json&prop=images&page=Polska`

Każdemu artykułowi towarzyszy oddzielna strona dyskusji[5], która służy do wymiany opinii, zgłaszania uwag i rozwiązywania konfliktów związanych z treścią artykułu. Z takich stron również zostały wykorzystane podobne zapytania podczas ekstrakcji miar.

Jeżeli chodzi o korzystanie z kopii zapasowych do ekstrakcji miar jakości, to można wymienić następujące pliki dla polskojęzycznej Wikipedii:

- **plwiki-latest-pages-meta-current.xml.bz2**: rekombinacja wszystkich stron (łącznie artykułów), tylko aktualne wersje. Ten plik służy do uzyskania większości miar artykułów.
- **plwiki-latest-pages-articles.xml.bz2**: zawiera artykuły, szablony, opisy obrazków / plików i podstawowe meta-strony. Może być również wykorzystany do uzyskania większości miar artykułów (z wyłączeniem miar dotyczących stron dyskusji).

[4] Wszystkie możliwe ustawienia serwisu API można znaleźć na specjalnej stronie: `https://pl.wikipedia.org/wiki/Specjalna:ApiSandbox`

[5] `https://pl.wikipedia.org/wiki/Pomoc:Strona_dyskusji`

- **plwiki-latest-pagelinks.sql.gz**: rekordy łącza stron wiki pomiędzy sobą. Używany do miar sieciowych - na przykład linki przychodzące z innych artykułów.
- **plwiki-latest-categorylinks.sql.gz**: rekordy opisujące przynależność stron wiki do kategorii. Może być stosowany do pomiaru liczby kategorii.
- **plwiki-latest-externallinks.sql.gz**: rekordy zewnętrzne adresów URL na stronach wiki. Może być użyty do pomiaru liczby linków wychodzących (zewnętrznych).
- **plwiki-latest-stub-meta-history.xml.gz**: zawiera tylko metadane z historii edycji stron wiki. Może być użyty do ekstrakcji liczby edytorów z różnych grup (botów, anonimowych użytkowników, administratorów itp.), oraz liczby edycji różnych typów (np. drobne zmiany, komentarze do edycji).
- **plwiki-latest-iwlinks.sql.gz**: informacja o linkach typu interwiki. Może być użyty do wyodrębnienia liczby unikalnych linków wewnętrznych (linki do innych artykułów Wikipedii).
- **plwiki-latest-templatelinks.sql.gz**: rekordy użycia szablonów przez strony wiki. Używane do określenia liczby szablonów oraz do sprawdzenia, czy artykuł posiada infoboks.
- **plwiki-latest-page.sql.gz**: dane bazowe o stronach wiki (identyfikator, tytuł, ograniczenia itp.). Może być użyty do wyodrębnienia czasu ostatniej edycji, długości strony w bajtach.
- **plwiki-latest-imagelinks.sql.gz**: rekordy użycia plików / obrazków na stronach wiki. Może być używany do zliczania obrazków.

Wyżej opisane pliki kopii zapasowych Wikipedii mogą dawać różne możliwości ekstrakcji miar. Na przykład w niektórych badaniach liczba obrazów jest brana pod uwagę w tagu [[image: ...]] w kodzie źródłowym strony wiki (Blumenstock, 2008b; Conti i in., 2014; Dalip i in., 2009; Stvilia i in., 2005a, 2005b; K. Wu, Zhu, Zhao i Zheng, 2010; Yaari, Baruchson-Arbib i Bar-Ilan, 2011; Zhang, Hu, Zhang i Yu, 2018). Jednak dodatkowe obrazy, umieszczone innym sposobem (na przykład za pomocą specjalnych szablonów), nie będą brane pod uwagę. Dlatego może być użyte inne podejście, które wyodrębniło liczbę obrazów z pliku rekordów użycia obrazów wiki (Lewoniewski, 2017b; Lewoniewski i Węcel, 2017; Lewoniewski, Węcel i Abramowicz, 2016, 2017b; Shang, 2018; Węcel i Lewoniewski, 2015).

Inny przykładem jest liczba linków wewnętrznych (interwiki) od strony i liczba przychodzących wewnętrznych linków do strony (z innych artykułów). W celu ekstrakcji tej miary można rozpatrywać kod źródłowy każdego artykułu, aby znaleźć linki do innych stron. Jednak w tym

przypadku nie mogą być zidentyfikowane linki, które zostały wstawione przez specjalne szablony. Dlatego może być używany również plik z rekordami łącza stron wiki pomiędzy sobą.

Niektórych miar nie można wyekstrahować z plików kopii zapasowych Wikipedii. Na przykład, aby uzyskać liczbę obserwatorów stron dla każdego artykułu, konieczne jest wysyłanie zapytania do Wikipedia API („English Wikipedia. API sandbox", nodate). Miary z zasobów zewnętrznych (takich jak Facebook, Twitter, Reddit itp.) należy również ekstrahować oddzielnie z innych źródeł.

5.3 Miary jakości artykułów Wikipedii

Na podstawie literatury (Anderka, 2013; Blumenstock, 2008a, 2008b; Dalip i in., 2009, 2011; Hu i in., 2007; Lewoniewski i Węcel, 2017; Lewoniewski i in., 2015, 2016; Lih, 2004; Lipka i Stein, 2010; Stvilia i in., 2005a; Warncke-wang i in., 2013; Wilkinson i Huberman, 2007a) oraz własnych badań w niniejszym rozdziale przedstawiono miary, które zostały następnie wzięte pod uwagę przy budowaniu modeli oceny jakości artykułów Wikipedii.

Modele te uwzględniają różne źródła, z których można pozyskać miary do oceny jakości informacji. Miary mogą również dotyczyć stron dyskusji nad artykułami.

Do modelu zostały wybrane miary różnego rodzaju, w tym ze stron dyskusji nad artykułami. Pełna lista miar artykułów wraz ze skrótami pokazana jest w tabeli 5.1.

Tabela 5.1. Miary jakości artykułów Wikipedii. Źródło: opracowanie własne

Skrót	Opis miary
A_1, A_2	Długość artykułu (w bajtach lub według liczby znaków)
A_3, A_4	Długość artykułu bez szumu (w bajtach lub według liczby znaków)
A_5	Liczba obrazków
A_6, A_7	Liczba obrazków z kodu wiki (w tekście lub w abstrakcie)
A_8	Liczba sekcji
A_9	Liczba sekcji z wyjątkiem źródłowych
A_{10}	Liczba kategorii, wpisanych do kodu
A_{11}, A_{12}	Liczba wszystkich szablonów (w tekście, w abstrakcie)
A_{13}	Liczba unikatowych szablonów
A_{14}, A_{15}	Liczba szablonów 1-go poziomu (w tekście lub w abstrakcie)
A_{16}, A_{17}	Długość szablonów 1-go poziomu (w tekście lub w abstrakcie)
A_{18}, A_{19}	Maksymalna długość szablonu (w tekście lub w abstrakcie)
A_{20}	Liczba szablonów o lukach jakości
A_{21}, A_{22}	Liczba linków wewnętrznych (w tekście lub w abstrakcie)
A_{23}, A_{24}	Długość tekstu bez referencji (w bajtach lub według liczby znaków)
A_{25}	Czas ostatniej zmiany artykułu

A_{26}, A_{27}	Długość abstraktu (w bajtach lub według liczby znaków)
A_{28}, A_{29}	Długość abstraktu bez szumu (w bajtach lub według liczby znaków)
A_{30}	Minimalna długość sekcji
A_{31}	Maksymalna długość sekcji
A_{32}	Liczba sekcji 1-go poziomu
A_{33}	Liczba sekcji 2-go poziomu
A_{34}	Liczba sekcji 1-go i 2-go poziomu (razem)
A_{35}	Liczba tabeli
A_{36-53}	Linki przychodzące z określonej przestrzeni nazw (ns0-ns3, ns6-ns15, ns100, ns101, ns828, ns829)
A_{54}	Linki przychodzące ze wszystkich rozpatrywanych przestrzeń nazw
A_{55}	Linki przychodzące z artykułów Wikipedii
A_{56}	Czas ostatniego odświeżenia
A_{57}	Czas ostatniego odświeżenia linków
A_{58}	Numer ostatniej edycji artykułu
A_{59}	Długość kodu źródłowego
A_{60}	Liczba przekierowań na stronę
A_{61}	Liczba wersji językowych
A_{62}	Data utworzenia artykułu
A_{63}	Liczba edycji artykułu
A_{64}	Liczba drobnych edycji
A_{65-68}	Liczba edycji w ciągu ostatnich 30, 90, 180, 365 dni
A_{69-72}	Liczba drobnych edycji w ciągu ostatnich 30, 90, 180, 365 dni
A_{73}	Liczba unikatowych autorów
A_{74}	Liczba unikatowych autorów anonimowych
A_{75-78}	Liczba unikatowych autorów dokonujących zmian w ciągu ostatnich 30, 90, 180, 365 dni
A_{79-82}	Liczba unikatowych autorów anonimowych dokonujących zmian w ciągu ostatnich 30, 90, 180, 365 dni
A_{83}	Liczba obserwujących
A_{84}	Suma odwiedzin za ostatni rok
A_{85}	Mediana odwiedzin w ciągu ostatnich 90 dni
A_{86}	Mediana odwiedzin w ciągu ostatnich 365 dni
A_{87}	Mediana odwiedzin w ciągu ostatnich 365 dni bez dni z brakiem odwiedzin
A_{88}, A_{89}	Liczba wszystkich referencji (w tekście lub w abstrakcie)
A_{90}, A_{91}	Liczba unikatowych referencji (w tekście lub w abstrakcie)
A_{92}	Gęstość referencji
A_{93}	Długość kodu referencji
A_{94}	Liczba referencji ze specjalnym szablonem
A_{95}	Średnia liczba wypełnionych parametrów w szablonach referencji
A_{96}	Liczba referencji posiadających archiwum
A_{97}	Liczba referencji posiadających co najmniej jeden specjalny identyfikator
A_{98-105}	Liczba referencji z określonym specjalnym identyfikatorem (DOI, ISBN, ISSN, OCLC, ARXIV, PMID, PMC, JSTOR)
A_{106}, A_{107}	Referencje z 50 oraz 100 najbardziej popularnych hostów w ramach wybranej wersji językowej Wikipedii
$A_{108-110}$	Referencje z 50, 100 oraz 300 nabardziej popularnych domen Internetu
$A_{111-115}$	Referencje z 50 najbardziej popularnych domen określonego państwa (Białoruś, Polska, Rosja, Ukraina, USA)

$A_{116-119}$	Referencje z szablonem określonego rodzaju (cytowania strony WWW, wiadomości, książki, czasopisma)
A_{120}, A_{121}	Długość strony dyskusji (w bajtach oraz według liczby znaków)
A_{122}	Liczba szablonów na stronie dyskusji
A_{123}	Liczba sekcji na stronie dyskusji
A_{124}	Liczba linków wewnętrznych na stronie dyskusji
A_{125}	Długość abstraktu na stronie dyskusji
A_{126}	Minimalna długość sekcji na stronie dyskusji
A_{127}	Maksymalna długość sekcji na stronie dyskusji
A_{128}	Czas ostatniego odświeżenia na stronie dyskusji
A_{129}	Czas ostatniego odświeżenia linków na stronie dyskusji
A_{130}	Numer ostatniej edycji na strony dyskusji
A_{131}	Długość kodu źródłowego strony dyskusji
A_{132}	Suma odwiedzin w ciągu ostatnich 365 dni strony dyskusji
A_{133}	Liczba sygnałów z serwisu Facebook do artykułu Wikipedii

Szczególnie warto opisać miary A_{36-53}, które związane z liczbą linków od stron Wikipedii z różnych grup. Strony w Wikipedii są pogrupowane w tzw. przestrzenie nazw (z ang. namespace czy ns). W zależności od wersji językowych liczba takich zdefiniowanych grup może się różnić. Na przykład, dla polskojęzycznej wersji zdefiniowano 29 przestrzeni nazw (Polska Wikipedia, 2018). Pomiędzy wersjami językowymi można znaleźć spójne oznaczenia przestrzeni nazw, które mogą okazać się przydatne do ekstrakcji miar jakości. Tabela 5.2 przedstawia oznaczenie oraz opis wybranych przestrzeni nazw stron Wikipedii. W dalszych analizach będą używane miary z wszystkich przedstawionych przestrzeń nazw za wyjątkiem ns4 oraz ns5, które zawierają informację ogólne na temat Wikipedii. Takie strony często zawierają linki do stron, które mają przypisane najwyższe oceny jakości - FA oraz GA (w niektórych wersjach językowych jeszcze SA). To jest związane m.in. z tym, że w przestrzeni nazw ns4 umieszczane są listy najlepszych artykułów. To oznacza, że artykuły z treścią podobnej jakości nie będą mieć linków z tych stron, co może prowadzić do zmniejszenia precyzji modeli jakości podczas ewaluacji nieocenionych artykułów.

Opis innych miar zostanie przedstawiony w następnych sekcjach. Dodatkowo w następnych sekcjach zostaną krótko opisane wymiary jakości z miarami, które do nich należą.

5.4 Miary jakości źródeł artykułów Wikipedii

Obszerna część niniejszej sekcji została opracowane na podstawie wcześniejszych badań (Lewoniewski, Węcel i Abramowicz, 2017a).

Tabela 5.2. Skróty oraz opisy wybranych przestrzeni nazw.

Oznaczenie	Opis przestrzeni nazw
ns0	Artykuły encyklopedyczne - główna przestrzeń nazw
ns1	Dyskusja artykułu
ns2	Wikipedyst(k)a - strony użytkowników Wikipedii, na podstronach – pomocnicze strony do własnego użytku, przyborniki, brudnopisy
ns3	Dyskusja wikipedysty/-ki
ns4	Wikipedia - informacje ogólne na temat Wikipedii, zasady oraz różnego rodzaju strony współpracy i dokumentacji
ns5	Dyskusja Wikipedii
ns6	Plik - strony opisujące załadowane pliki multimedialne (obraz, dźwięk, film)s
ns7	Dyskusja pliku
ns8	MediaWiki - komunikaty interfejsu.
ns9	Dyskusja MediaWiki
ns10	Szablon - strony przeznaczone do definiowania oraz opisywania szablonów
ns11	Dyskusja szablonu
ns100	Portal - strony prezentujące czytelnikom w atrakcyjny sposób tematyczny wycinek zawartości Wikipedii
ns101	Dyskusja portalu
ns828	Moduł - strony zawierające kod w języku Lua, współpracujące ze skomplikowanymi szablonami
ns829	Dyskusja modułu

Źródło: Opracowane na podstawie (Polska Wikipedia, 2018)

Artykuły z Wikipedii o wysokiej jakości muszą być dobrze zbadane i mieć reprezentatywny przegląd odpowiedniej literatury. Podczas dodawania lub edytowania treści artykułu autorzy muszą dodawać wiarygodne i ogólnodostępne źródła. W ten sposób osoby korzystające z tej encyklopedii mogą sprawdzić, skąd pochodzą informacje i zweryfikować opisane w nich fakty.

Istnieją badania, które skupiają się na analizie jakościowej referencji w Wikipedii. Jedno z pierwszych badań w tym kierunku sugerowało, że artykuły Wikipedii mają tendencję do cytowania artykułów w czasopismach o dużym wpływie, takich jak „New England Journal of Medicine", „Nature", „Science" (Nielsen, 2007). Jednocześnie rośnie liczba recenzowanych prac naukowych z nauk o zdrowiu, które cytują artykuły Wikipedii (Bould i in., 2014). Referencje mogą obejmować szeroki zakres tematów, ale w szczególności dotyczy to artykułów z ekologii, ewolucji i innych tematów, które mogą wzbogacić encyklopedię o źródła naukowe (Lin i Fenner, 2014). Ponad połowa referencji używanych w artykułach historycznych encyklopedii to źródła internetowe, takie jak wiadomości, media, strony rządowe (Luyt i Tan, 2010). Jeśli użytkownicy dodają odwołania związane z publikacjami naukowymi, wolą używać bardziej książki niż artykuły naukowe jako źródła (Kousha i Thelwall, 2017). Wikipedia jest szczególnie cenna ze względu na potencjalne bezpośrednie powiązania z innymi źródłami pierwotnymi za pomocą specjalnego identyfikatora, takiego jak DOI lub PubMed ID (R. D. Page, 2010). Dodatkowo akademicki status pracy jest najważniejszym predyktorem jego pojawienia się w odnośnikach Wikipedii (Teplitskiy, Lu i Duede, 2017).

Społeczność użytkowników Wikipedii opracowała również zestaw szablonów do zgłaszania artykułów, które nie mają wystarczającej liczby referencji lub nie ma żadnych wskazanych źró-

deł. W anglojęzycznej wersji Wikipedii takie szablony są najczęściej używane spośród ponad 300 szablonów wskazujących na różne wady jakości (Anderka, 2013). Zatem można stwierdzić, że społeczność Wikipedii zwraca szczególną uwagę na dostępność referencji w artykułach.

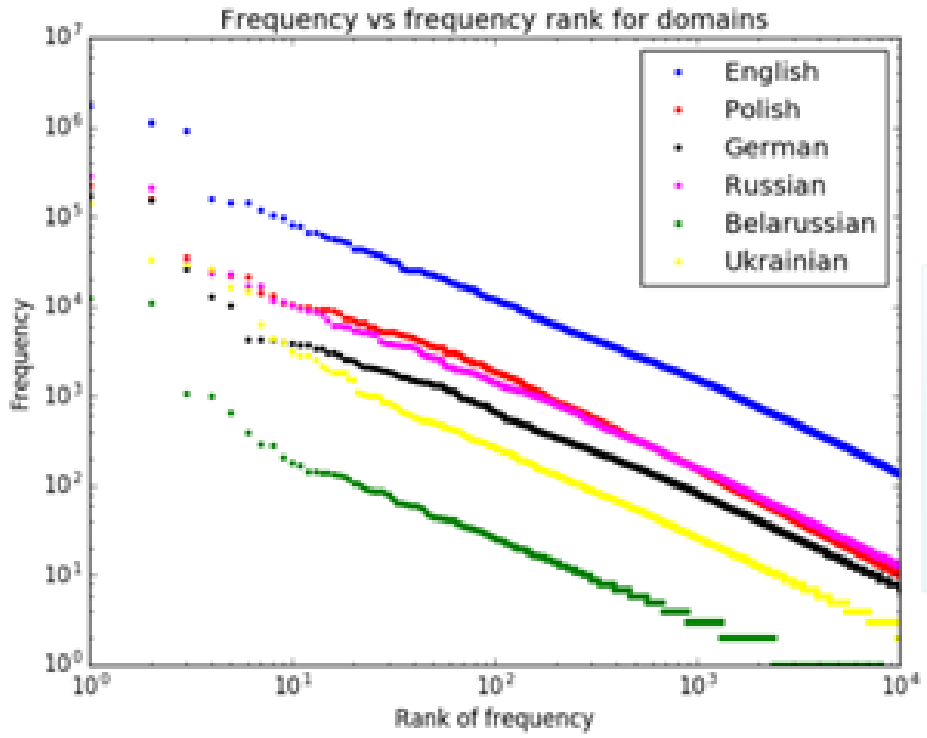
W kodzie źródłowym stron wiki odwołania są zwykle umieszczane pomiędzy specjalnymi znacznikami *<ref>...</ref>*. Ogólnie możemy podzielić te referencje na dwie grupy: ze specjalnym szablonem i bez niego. Referencje bez specjalnego szablonu zazwyczaj mają adres URL źródła i opcjonalny opis (np. tytuł). Referencje ze specjalnymi szablonami mogą mieć różne dane opisujące źródło. Tutaj w osobnych polach możemy dodać informacje o autorze (autorach), tytule, adresie URL, formacie, dacie dostępu, wydawcy i innych. Ponadto szablony te mogą zawierać specjalne identyfikatory, takie jak DOI, JSTOR, PMC, PMID, arXiv, ISBN, ISSN i OCLC. Zestaw możliwych parametrów zależy od rodzaju szablonów, które mogą opisywać źródło sieciowe, książkę, czasopismo, wiadomości, konferencję, akt prawny i inne. Dodatkowo, każda wersja językowa Wikipedii może korzystać z własnej grupy szablonów o własnych nazwach i zestawu parametrów opisujących źródła informacji.

W celu ekstrakcji informacji o źródłach został utworzony specjalny algorytm (Lewoniewski, Węcel i Abramowicz, 2017a), który bierze pod uwagę różne nazwy szablonów referencji i zestawy parametrów z różnych wersji językowych Wikipedii. Na przykład, przy badaniu 7 wersji jezykowych (BE, DE, EN, FR, PL, RU, UK), mających około 10 mln artykułów, można odtrzymać informację o ponad 42 mln referencji z ponad 3 mln domen internetowych. Dystrybucja częstotliwości domen źródeł w każdym języku pokazana jest na rysunku 5.3.

5.4.1 Unifikacja danych referencji w różnych wersjach językowych Wikipedii

Specjalne identyfikatory mogą pomóc w znalezieniu torzsamych referencj, nawet w przypadku różnych parametrów w opisie (np. tytuły w innych językach). Możemy również ujednolicić ich adres URL. Na przykład, jeśli numer referencyjny to ISBN „978-3-319-46254-7”, podajemy adres URL ”books.google.com/books?vid=ISBN9783319462547”. Bardziej szczegółowe informacje na temat identyfikatorów, które były wykorzystane do ujednolicenia odniesień, przedstawiono w tabeli 5.3.

Tabela 5.4 przedstawia liczbę unikalnych referencji z konkretnym identyfikatorem w każdej z 5 wersji językowych Wikipedii.



Rysunek 5.3. Dystrybucja częstotliwości domen źródeł Wikipedii a każdym z 7 wersji językowych.

Źródło: (Lewoniewski, Węcel i Abramowicz, 2017a)

Tabela 5.3. Identyfikatory używane do unifikacji referencji Wikipedii

Ident.	**Opis**	**Nowy URL**
arXiv	arXiv repository identifier	*http://arxiv.org/abs/...*
DOI	Digital object identifier	*http://doi.org/...*
ISBN	International Standard Book Number	*http://books.google.com/books?vid=ISBN...*
ISSN	International Standard Serial Number	*https://worldcat.org/ISSN/...*
JSTOR	Journal Storage number	*https://jstor.org/stable/...*
PMC	PubMed Central	*https://ncbi.nlm.nih.gov/pmc/articles/PMC...*
PMID	PubMed	*https://ncbi.nlm.nih.gov/pubmed/...*
OCLC	WorldCat's Online Computer Library Center	https://worldcat.org/oclc/...

Źródło: (Lewoniewski, Węcel i Abramowicz, 2017a)

Tabela 5.4. Liczba referencji z konkretnym identyfikatorem w artykułach Wikipedii.

lang.	arXiv	DOI	ISBN	ISSN	JSTOR	PMC	PMID	OCLC
BE	90	1,185	13,656	78	28	53	198	19
EN	4,226	1,014,602	1,670,495	79,442	35,709	16,384	52,387	54,995
PL	577	41,796	245,833	23,319	781	338	11,157	1,131
RU	1,577	33,956	232,427	3,045	785	1,236	5,164	977
UK	301	2,562	37,628	618	96	160	313	400
Total	10,029	1,175,496	27,03,705	144,495	41,035	19,975	80,106	67,791

Źródło: (Lewoniewski, Węcel i Abramowicz, 2017a)

5.4.2 Podobieństwo referencji

Ujednolicenie adresów URL na podstawie identyfikatorów pozwoliło na zliczenie unikatowych referencji i może być użyte do porównania podobieństwa referencji w artykułach Wikipedii w różnych wersjach językowych.

Przy rozpatrywaniu wszystkich artykułów w 7 wersjach językowych, około 6,9 miliona z nich mają co najmniej 1 referencję (Lewoniewski, Węcel i Abramowicz, 2017a). Po ekstrakcji danych z artykułów otzymano dane dla około 30 mln unikatowych referencji. Tabela 5.5 przedstawia wyniki porównania liczby wspólnych referencji w badanych wersjach językowych Wikipedii.

Tabela 5.5. Liczba wspólnych referencji użytych w wersjach językowych Wikipedii.

lang.	BE	DE	EN	FR	PL	RU	UK
BE	**82,295**	3,522	19,116	6,127	5,043	47,931	13,100
DE	-	**2,988,443**	345,202	81,572	41,558	69,634	21,097
EN	-	-	**18,470,130**	584,037	244,120	635,546	160,408
FR	-	-	-	**3,364,409**	61,104	118,700	32,470
PL	-	-	-	-	**1,548,696**	71,221	26,022
RU	-	-	-	-	-	2,873,070	185,473
UK	-	-	-	-	-	-	**635,149**

Źródło: Obliczenia własne w maju 2017r.

Największą liczbę referencji w angielskiej Wikipedii można wytłumaczyć największą liczbą artykułów w niej zawartych. Najwięcej wspólnych referencji ta wersja językowa ma z rosyjską Wikipedią.

Warto zaznaczyć, iż polskojęzyczna wersja ma więcej wspólnych referencji z rosyjską oraz ukraińską wersją (RU i UK odpowiednio) niż z niemiecką i francuską (DE i FR odpowiednio). Z drugiej strony, francuska Wikipedia ma znacznie więcej wspólnych referencji niż razem polska, ukraińska oraz białoruska z angielską. Opisane różnice mogą m.in. pokazywać jaka część informacji w artykułach Wikipedii mogą być spójne pomiędzy wersjami językowymi.

Znajomość adresów URL wszystkich referencji daje możliwość do identyfikowania najbardziej popularnych domen. Tabela 5.6 pokazuje 10 najbardziej popularnych witryn internetowych, z których pochodzą referencje w każdej wersji językowej Wikipedii.

Tabela 5.6. 10 najpopularniejszych domen referencji w różnych wersjach językowych Wikipedii.

BE	EN	PL
books.google.com	books.google.com	books.google.com
pravo.by	books.google.de	web.archive.org
football.by	spiegel.de	doi.org
doi.org	doi.org	sports-reference.com
cuetracker.net	welt.de	archive.is
naviny.org	zeit.de	worldcat.org
by.tribuna.com	faz.net	stat.gov.pl
worldsnooker.com	worldcat.org	discogs.com
web.archive.org	youtube.com	allmusic.com
gks.ru	sueddeutsche.de	getamap.ordnancesurvey.co.uk

RU	UK
books.google.com	insee.fr
doi.org	books.google.com
insee.fr	kia.hu
billboard.com	w1.c1.rada.gov.ua
textual.ru	demo.istat.it
int.soccerway.com	nsi.bg
lenta.ru	cvk.gov.ua
web.archive.org	pravda.com.ua
youtube.com	youtube.com
kommersant.ru	web.archive.org

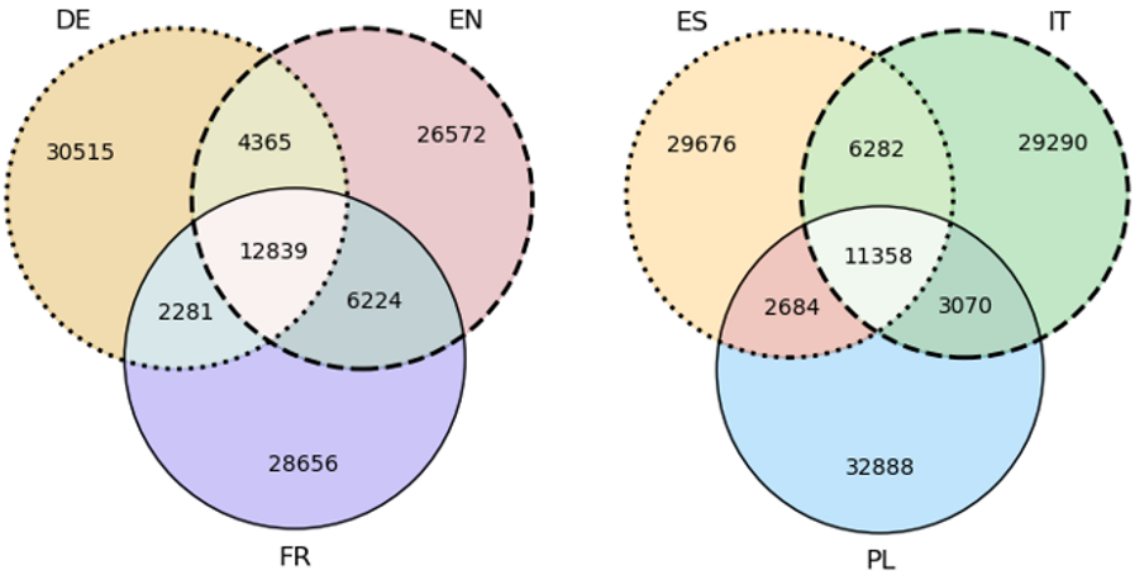
Źródło: Obliczenia własne.

Rysunek 5.4 pokazuje pokrycie najpopularniejszych 50 tys. domen w referencjach w wybranych wersjach językowych Wikipedii.

5.5 Miary SEO

Obszerna część niniejszej sekcji została opracowane na podstawie wcześniejszych badań (Lewoniewski, Härting, Węcel, Reichstein i Abramowicz, 2018).

W badanach z 2017-2018 roku stwierdzono, że ponad 4 mld osób ma dostęp do usług internetowych i baz danych (Internet World Stats, 2018). Liczba ta stanowi 50% światowej populacji. W przyszłości spodziewany jest trwający trend wzrostowy (International Telecommunica-



Rysunek 5.4. Pokrycie 50 tys. najpopularniejszych domen w referencjach w wybranych wersjach językowych Wikipedii.

Źródło: (Lewoniewski, Härting, Węcel, Reichstein i Abramowicz, 2018)

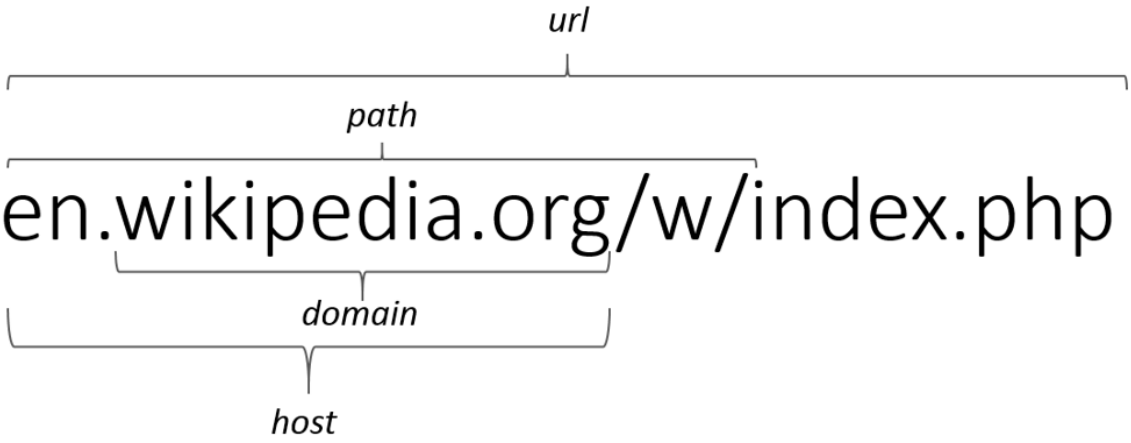
tion Union, 2017). W oparciu o ten rozwój wzrasta użycie narzędzi Search Engine Optimization (SEO). SEO to narzędzia i metody optymalizacji stron internetowych pod wyszukiwarki.

Jedną z części SEO jest planowanie i optymalizacja projektów internetowych. Jednak SEO oznacza również monitorowanie wydajności strony internetowej. W tym kontekście można rozróżnić podejścia do kontrolowania wydajności obecności w Internecie (Kumar i Kumar, 2014; Schroeder, 2007).

Narzędzia, które sprawdzają pojedyncze wskaźniki wydajności, a zwłaszcza stopień widoczności stron internetowych, stanowią bardzo skuteczne podejście w obszarze SEO. W marketingu cyfrowym narzędzia SEO są bardzo rozpowszechnionym segmentem aplikacji. Takie narzędzia obliczają miary wyszukiwarek jako wskaźniki KPI (z ang. Key Performance Indicators). W tym kontekście istnieje kilka przedsiębiorstw, które oferują szeroką gamę konkretnych narzędzi. Jednym z najpopularniejszych narzędzi wśród europejskiej społeczności SEO jest SISTRIX (Härting, Mohl, Steinhauser i Möhring, 2016; SISTRIX GmbH, 2018a).

SISTRIX jest przede wszystkim narzędziem analitycznym. Korzystanie z zestawu narzędzi zapewnia propozycje ulepszeń własnych stron oraz umożliwia pomiar i monitorowanie aspektów jakościowych projektów internetowych. Zasadniczo warto monitorować różne strony równolegle i analizować zmiany stron internetowych w czasie (Drèze i Zufryden, 2004).

Oprócz wspólnych wskaźników wydajności dla analizy wyszukiwarek, takich jak linki zwrotne i profil słów kluczowych, SISTRIX oferuje również bardziej zaawansowane miary, takie jak wskaźnik widoczności oraz miary społeczne. SISTRIX pozwala na przeglądanie miar strony z różnych poziomów: domeny, hosta, ścieżki, adresu URL. Rysunek 5.5 pokazuje przykład strony internetowej z Wikipedii i poziomów, które można oddzielnie analizować.

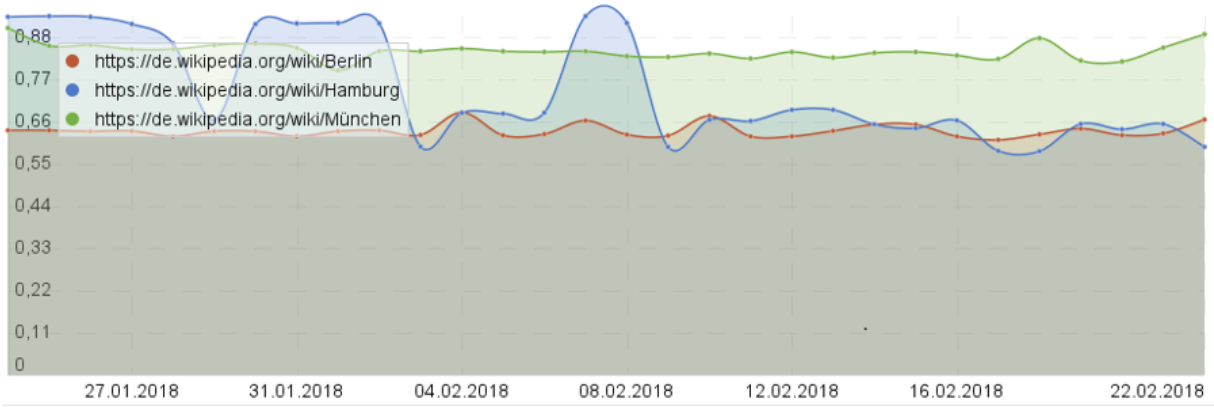


Rysunek 5.5. Domena, host, ścieżka i adres URL strony na przykładzie Wikipedii

Źródło: Opracowanie własne.

Niżej zostały opisane miary z narzędzia SISTRIX, które będą używane w dalszej analizie.

Wskaźnik widoczności (Visibility Index) jest dobrze znaną i rozpowszechnioną miarą w narzędziach SEO (Goodman, 2008). Jeśli chodzi o SISTRIX, kluczowy wskaźnik wydajności (KPI) jest obliczany na podstawie dedykowanej bazy danych i pokazuje widoczność domeny lub adresu URL stron w wynikach wyszukiwania w Google (SISTRIX GmbH, 2018a). Wkaźnik widoczności domeny lub adresu URL jest zwykle generowany przez pulę słów kluczowych. Słowa kluczowe są klasyfikowane i ważone w wynikach wyszukiwania Google (Maynes i Everdell, 2014). Każdego tygodnia SISTRIX oblicza wskaźnik za pomocą puli słów kluczowych obejmujących milion słów oraz kombinacji słów. 10% tych słów kluczowych jest tworzonych przez bieżące wydarzenia. 90% zawsze pozostaje bez zmian. 100 najlepszych pozycji w Google jest analizowanych i rejestrowanych co tydzień. Wyniki są ważone zgodnie z aspektami pozycji i oczekiwanej liczby wyszukiwań dla każdego słowa kluczowego (SISTRIX GmbH, 2018b). Podobnie jak Google, SISTRIX utrzymuje zmienny algorytm obliczania takiego indeksu widoczności (Härting i in., 2016). Na rysunku 5.6 pokazany przykład wyników mierzenia wskaźnika widoczności dla wybranych artykułów Wikipedii w zaznaczonym terminie.



Rysunek 5.6. Porównanie wskaźników widoczności dla artykułów Wikipedii.

Źródło: (SISTRIX GmbH, 2018a)

Znaczenie wskaźnika widoczności dla działań związanych z SEO otrzymuje różne oceny (Härting i in., 2016). Po pierwsze, wskaźnik widoczności nie jest mierzony w czasie rzeczywistym i nie może dostarczyć danych o ruchu organicznym na stronie internetowej. Po drugie, witryny niszowe z dużą liczbą specjalnych słów kluczowych będą automatycznie miały niższy wskaźnik widoczności. Nawiązując do badań w tym kierunku (Härting i in., 2016), można znaleźć korelację między wskaźnikiem widoczności a jakością strony internetowej, ponieważ zachowanie algorytmu Google jest istotnym czynnikiem wskaźnika widoczności SISTRIX. Wyszukiwarka Google ustawia wyższy priorytet dla witryn o wysokiej jakości treści i trafności. Dlatego jakość może mieć wpływ na wskaźnik widoczności. Problem polega na tym, że witryny skupiające się na niszowych treściach i słowach kluczowych nie są czasami umieszczane w bazie narzędzi SEO, nawet jeśli są dobrej jakości (RYTE GmbH, 2018).

Wcześniejsze badania pokazały, że bardziej popularne artykuły Wikipedii zazwyczaj mają wyższy wskaźnik widoczności (Lewoniewski, Härting i in., 2018). Wskaźnik widoczności w SISTRIX może być badany z punktu widzenia użytkowników różnych państw. Badania pokazują, że referencje od lokalnych hostów częściej mają niezerowy wskaźnik widoczności (Lewoniewski, Härting i in., 2018). Np. referencje pochodzące z polskich stron internetowych zazwyczaj mają wyższy wskaźnik widoczności dla użytkownika z Polski. Dotyczy to również samych artykułów Wikipedii w języku polskim.

Liczba oraz profil słów kluczowych (LPSK) pokazuje, ile różnych słów kluczowych domena (host lub adres URL) osiągnęła w rankingu Top-100 (SISTRIX GmbH, 2018a). W narzędziu SISTRIX ta miara ma nazwę „SEO". Rozszerzone uwzględnienie dystrybucji słów kluczowych prowadzi do utworzenia profilu słów kluczowych. Profil słów kluczowych jest częścią sekcji SEO Overview

w SISTRIX Toolbox. Umożliwia głębsze zrozumienie struktury słów kluczowych na podstawie oceny ilościowej. W SEO słowa kluczowe są ważnym czynnikiem w zakresie optymalizacji na stronie. Szczególnie pozycja słowa kluczowego (np. w tagu H1) przyczynia się w dużym stopniu do rankingu Google. Oznacza to, że Google klasyfikuje strony poświęcone konkretnemu tematowi lub słowu kluczowemu jako bardziej odpowiednie. Pod tym względem pozycja w wynikach Google, mierzona przez profil słów kluczowych SISTRIX, może wskazywać na istotność strony. W tym kontekście można założyć, że wyniki profili słów kluczowych SISTRIX również wykazują wyraźną gradację jakości artykułów Wikipedii.

Linki zwrotne (z angl. *backlinks*) są to link zewnętrzne, które odwołują się do rozpatrywanej domeny (SEO Glossary, 2018). Liczba linków zwrotnych liczy liczbę odsyłaczy z określonej strony internetowej wyświetlanej w sieci WWW (Klusch, 2001). Zanim wyszukiwarki takie jak Google i Yahoo zyskały tak ogromne znaczenie, linki zwrotne stały się jednym z głównych środków nawigacji internetowej. Wraz z nadejściem wyszukiwarek liczba linków zwrotnych stała się głównym czynnikiem analizy popularności i znaczenia strony internetowej (Killoran, 2013). W dzisiejszych czasach linki zwrotne są nadal ważne dla SEO, ale koncentrują się bardziej na badaniu jakości łącza, a nie na jego ilości (RYTE GmbH, 2018). Można przypuścić, że linki zwrotne wysokiej jakości wpływają na jakość artykułów Wikipedii, ponieważ z jednej strony Google (jako najważniejsza wyszukiwarka na świecie) nadaje duże znaczenie tym linkom, a drugiej strony linki do zaufanych domen są zazwyczaj dobrym wskaźnikiem jakości.

Sygnały społecznościowe są to odnośniki do badanych stron z serwisów społecznościowych (takich jak Facebook, Twitter i innych). Od powstania Facebooka lub innych sieci społecznościowych sygnały społecznościowe stały się również ważną informacją dla analizy SEO. Te miary dostarczają informacji na temat interakcji społecznej, zachowań społecznych i relacji społecznych (RYTE GmbH, 2018). Sygnały społecznościowe mogą na przykład zawierać komentarze, polubienia lub udostępnienia na Facebooku. W SISTRIX, wskaźnik sygnału społecznego mierzy typ i ilość sygnałów społecznościowych dostępnych dla URL czy domeny. SISTRIX obejmuje pięć dużych sieci społecznościowych: Facebook, Twitter, Google+, LinkedIn i Pinterest. Korelacja między sygnałem społecznym a pozycją w rankingu adresu URL jest wyjątkowo wysoka. Dotyczy to wszystkich sieci społecznościowych objętych SISTRIX i prowadzi do założenia, że wyniki sygnałów społecznościowych są również skorelowane z jakością artykułów Wikipedii. W niniejszej pracy obliczone sygnały z serwisu społecznościowego Facebook do poszczególnych artykułów Wikipedii w różnych wersjach językowych to obliczenia miary A133.

Badania pokazują, że podobnie jak i w przypadku wskaźnika widoczności, popularne artykuły Wikipedii mają wyższe wartości sygnałów społecznościowych. Dodatkowo, znacznie większa część sygnałów społecznościowych przypada na artykuły wysokiej jakości (Lewoniewski, Härting i in., 2018).

5.6 Wymiary jakości artykułów Wikipedii

W tej sekcji został przedstawiony autorski podział miar jakości na poszczególne wymiary: aktualność, czytelność, kompletność, obiektywność, relewantność, styl, wiarygodność. Każdy wymiar opisany został w osobnej podsekcji. Dodatkowo została pokazana wielowymiarowość niektórych miar. Wymiary zostały wybrane na podstawie analizy literatury oraz własnych obserwacji (patrz. rozdział „Metody określenia jakości artykułów Wikipedia"). Dodatkowo został wprowadzony nowy wymiar jakości dla miar, określających popyt na informacje przedstawioną w artykułach.

5.6.1 Aktualność

Informacje na niektóre tematy mogą się zmieniać wraz z upływem czasu (osoby żyjące, zamieszkałe miejsca itp.). Dlatego ważne jest, aby artykuł zawierał dane zgodne z aktualnym stanem rzeczywistości. Niektóre miary mogą pomóc w ocenie tego aspektu jakości: liczba unikalnych redaktorów i liczba edycji (Arazy, 2010; Jemielniak, 2013; Suzuki i Yoshikawa, 2012; Wilkinson i Huberman, 2007a) za wybrany czas.

Do tego wymiaru należą następujące miary jakości: A25, A56-A58, A65-A72, A75-A78, A79-A82, A128-A130.

5.6.2 Czytelność

Miary związane z tym wymiarem jakości muszą pokazywać, w jakim stopniu tekst jest zrozumiały i wolny od niepotrzebnej złożoności. Dlatego przede wszystkim należy wziąć pod uwagę specjalne formuły czytelności, takie jak Zautomatyzowany Wskaźnik Czytelności (Automated Readability Index) (Blumenstock, 2008a; Dalip i in., 2011; Dang i Ignat, 2016a; Flekova, Ferschke i Gurevych, 2014; Liu i Ram, 2018; Ransbotham i Kane, 2011; Ransbotham, Kane i Lurie, 2012; Senter i Smith, 1967; Shen, Qi i Baldwin, 2017), Wskaźnik Bormutha (Bormuth Index)

(Anderka, 2013; Bormuth, 1966), wskaźnik Coleman-Liau (Blumenstock, 2008a; Coleman i Liau, 1975; Dalip i in., 2011; Dang i Ignat, 2016a; Flekova i in., 2014; Ransbotham i Kane, 2011; Shen i in., 2017), czytelność FORCAST (Blumenstock, 2008a; Caylor i Sticht, 1973), wskaźnik Flescha (Blumenstock, 2008a; Conti i in., 2014; Dalip i in., 2011; Dang i Ignat, 2016a; di Sciascio, Strohmaier, Errecalde i Veas, 2017; Flekova i in., 2014; Flesch, 1948; Shen i in., 2017; Stvilia i in., 2005a, 2005b; K. Wu i in., 2010), poziom jakości Flesch-Kincaid (Blumenstock, 2008a; Conti i in., 2014; Dalip i in., 2011; Dang i Ignat, 2016a; di Sciascio i in., 2017; Flekova i in., 2014; Kincaid, Fishburne Jr, Rogers i Chissom, 1975; Shen i in., 2017; Stvilia i in., 2005a, 2005b; Warncke-wang i in., 2013; K. Wu i in., 2010), wskaźnik Gunning Fog (Blumenstock, 2008a; Dalip i in., 2011; Dang i Ignat, 2016a; Flekova i in., 2014; Gunning, 1952; Shen i in., 2017), LIX (Dalip i in., 2011; Flekova i in., 2014), wskaźnik czytelności Miyazaki EFL (Anderka, 2013; Greenfield, 1999), Dale-Chall (Dale i Chall, 1948; Dang i Ignat, 2016a; Shen i in., 2017), Cieniowanie SMOG (Blumenstock, 2008a; Dang i Ignat, 2016a; Flekova i in., 2014; Mc Laughlin, 1969; Shen i in., 2017), wzór Linseara (Chen, 2012; Dang i Ignat, 2016a; Shen i in., 2017) oraz inne. Wzory te często opierają się na wcześniej wyliczonych słowach różnych typów. Zatem ten wymiar może również składać się z różnych miar językowych. W zależności od wersji językowej można zdefiniować ponad 100 takich miar (Lewoniewski, Khairova, Węcel, Stratiienko i Abramowicz, 2017; Lewoniewski, Wecel i Abramowicz, 2017)

Wyjaśnienia może wymagać również pojęcie *szumu* – są to wszelkie znaki, które nie niosą treści, ale są wykorzystywane do np. formatowania tekstu czy tworzenia szablonów.

Również tutaj wymienione są miary związane z charakterystyką poszczególnych słów w tekście. Ekstrakcja większości z wymienionych miar wiąże się z bardziej złożoną analizą, która wymaga dodatkowych predefiniowanych zasobów (słowników) dla każdej wersji językowej. Niestety większość dostępnych miar dotyczy języka angielskiego, a opracowanie ich dla innych języków nie jest trywialne.

W trakcie eksperymentów zostały wyeliminowane te miary, które nie mogą być stosowane w ramach każdej z rozpatrywanych wersji językowych - głównie to miary lingwistyczne. W związku z tym, do wymiaru czytelności zostały przypisane następujące miary jakości: A1-A4, A8-A24, A26-A35.

5.6.3 Kompletność

Artykuły w Wikipedii o wysokiej jakości muszą dbać o umieszczenie istotnych faktów i szczegółów w artykułach Wikipedia. Miary z tego wymiaru pochodzą z tekstu artykułu i dotyczą głównie liczby znaków i słów. Mamy tutaj przede wszystkim statystyki tekstowe. Jedną z najpopularniejszych miar tego wymiaru jest objętość treści mierzona długością artykułów (Blumenstock, 2008a; Conti i in., 2014; Dalip i in., 2011; Dang i Ignat, 2016a; Flekova i in., 2014; Lerner i Lomi, 2018; Lewoniewski, 2017b; Lewoniewski i Węcel, 2017; Lewoniewski i in., 2016, 2017b; Ransbotham i Kane, 2011; Shen i in., 2017; Stvilia i in., 2005a, 2005b; Warncke-Wang i in., 2015; Węcel i Lewoniewski, 2015; K. Wu i in., 2010; Yaari i in., 2011; Zhang i in., 2018). Długość można mierzyć na różne sposoby: bajty, znaki, słowa czy liczebność.

Do tego wymiaru należą następujące miary jakości: A1-A10, A16-A19, A21-24, A26-A35, A59, A93-95, A122-A127.

5.6.4 Obiektywność

Artykuły w Wikipedii muszą przedstawiać informacje w sposób rzetelny i bezstronny. Artykuły tworzone przez większą liczbę osób mogą być bardziej obiektywne, stąd jedną z miar może być liczba unikalnych autorów (Conti i in., 2014; di Sciascio i in., 2017; Ferschke, Gurevych i Rittberger, 2012; Flekova i in., 2014; Jemielniak, 2013; Kane, 2011; Kittur i Kraut, 2008; Lewoniewski i in., 2016; Lih, 2004; Liu i Ram, 2018; Stvilia i in., 2005a, 2005b; Węcel i Lewoniewski, 2015; Wilkinson i Huberman, 2007a, 2007b; K. Wu i in., 2010; Yaari i in., 2011). Tutaj mogą być również używane miary związane z badaniem liczby obrazów (Blumenstock, 2008b; Conti i in., 2014; Dalip i in., 2009; Kane, 2011; Lewoniewski, 2017b; Lewoniewski i Węcel, 2017; Lewoniewski i in., 2016, 2017b; Liu i Ram, 2018; Shang, 2018; Stvilia i in., 2005a, 2005b; Węcel i Lewoniewski, 2015; K. Wu i in., 2010; Yaari i in., 2011; Zhang i in., 2018) Ten wymiar zawiera również miary dotyczące stron dyskusji artykułów.

Do tego wymiaru należą następujące miary jakości: A5-A7, A63, A65-83, A120-127, A131.

5.6.5 Popyt

Do tego wymiaru należą głównie miary określające popularność artykułu i popytu na informacje zawarte w nim ze strony czytelników oraz autorów Wikipedii. Popularność może odgrywać ważną rolę w szacowaniu jakości w konkretnych wersjach językowych Wikipedii (Lewoniewski

i in., 2016). Większa liczba użytkowników czytających artykuł może wpływać na szybkość zauważania oraz poprawiania błędów, częściej mogą być wprowadzane zmiany (w tym w celu aktualizacji danych).

Popularność artykułu można mierzyć na podstawie liczby odwiedzin strony (Lewoniewski i in., 2016, 2017b). W związku z tym, do tego wymiaru należą następujące miary jakości: A84-A87.

5.6.6 Relewancja

Ten wymiar pokazuje, na ile istotne dla autorów oraz czytelników są wybrane artykuły Wikipedii. Z tego powodu można używać takich miar jak wiek artykułu (Conti i in., 2014; Dalip i in., 2009; di Sciascio i in., 2017; Ferschke i in., 2012; Flekova i in., 2014; Kane, 2011; Kittur i Kraut, 2008; Lerner i Lomi, 2018; Lewoniewski, 2017b; Lewoniewski i in., 2016; Liu i Ram, 2018; Ransbotham i Kane, 2011; Ransbotham i in., 2012; Stvilia i in., 2005a, 2005b; Warncke-wang i in., 2013; K. Wu i in., 2010), liczba obserwatorów stron (Lewoniewski i in., 2016; Węcel i Lewoniewski, 2015), liczba linków przychodzących wewnętrznych (ile razy artykuł jest cytowany przez inne artykuły Wikipedii) (Dalip i in., 2009; Ferschke i in., 2012; Flekova i in., 2014; Lewoniewski i in., 2016; Soonthornphisaj i Paengporn, 2017; Węcel i Lewoniewski, 2015) oraz inne, w tym bardziej złożone (np. PageRank (L. Page, Brin, Motwani i Winograd, 1999)). Można również wziąć pod uwagę miary, które pokazują liczbę linków do artykułów Wikipedii ze źródeł zewnętrznych, takich jak Reddit (Moyer, Carson, Dye, Carson i Goldbaum, 2015), Facebook, Twitter oraz innych serwisów społecznościowych (Lewoniewski, Härting i in., 2018).

Do tego wymiaru należą następujące miary jakości: A21, A22, A54, A55, A60-87, A124, A132-A133.

5.6.7 Styl

Artykuły w Wikipedii o wysokiej jakości muszą być przygotowane zgodne ze wskazówkami dotyczącymi stylu, który dotyczy m.in. organizacji oraz struktury artykułu. Zatem jedną z najprostszych i najpopularniejszych miar tego wymiaru jest liczba sekcji w artykule (Blumenstock, 2008b; Conti i in., 2014; Dalip i in., 2009; Ferschke i in., 2012; Lerner i Lomi, 2018; Lewoniewski i in., 2016; Warncke-wang i in., 2013; Węcel i Lewoniewski, 2015; K. Wu i in., 2010; Zhang i in., 2018). Tutaj również mogą być użyte takie miary, jak liczba tabel (Anderka, 2013; Blumenstock,

2008b), liczba szablonów (Anderka, 2013; Lerner i Lomi, 2018; Lewoniewski, 2017b; Warncke-wang i in., 2013). Na styl mogą wplywać również użycie specjalnych wiki-znaczników, czy też szablonów {{...}}.

Część miar z tego wymiaru związana również z analizą części mowy. Ilekroć jest mowa o wskaźniku, chodzi o udział procentowy określonego zjawiska, np. wskaźnik pytań wyznaczamy poprzez podzielenie liczby pytań przez liczbę zdań. Część miar związana jest ze strukturą artykułów.

Do tego wymiaru należą następujące miary jakości: A5-A7, A11-A20, A26-A31, A35, A59, A92, A94.

5.6.8 Wiarygodność

Używanie wiarygodnych źródeł w Wikipedii jest jednym z ważnych kryteriów pisania artykułów o wysokiej jakości („Wikipedia - Featured article criteria", nodate). Czytelnicy encyklopedii muszą mieć możliwość sprawdzenia, skąd pochodzą informacje („Wikipedia - Verifiability", nodate). Dlatego jedną z najczęściej używanych miar związanych z wiarygodnością jest liczba referencji w artykułach Wikipedii (Blumenstock, 2008b; Conti i in., 2014; Dalip i in., 2009; Dang i Ignat, 2016a; di Sciascio i in., 2017; Ferschke i in., 2012; Lewoniewski, 2017b; Lewoniewski i Węcel, 2017; Lewoniewski i in., 2016, 2017b; Shen i in., 2017; Soonthornphisaj i Paengporn, 2017; Węcel i Lewoniewski, 2015; K. Wu i in., 2010; Zhang i in., 2018) lub linki zewnętrzne (Blumenstock, 2008b; Conti i in., 2014; Dalip i in., 2009; Ferschke i in., 2012; Flekova i in., 2014; Lewoniewski i in., 2016; Soonthornphisaj i Paengporn, 2017; Stvilia i in., 2005a; Węcel i Lewoniewski, 2015; Yaari i in., 2011). Z pokrewnych badań wynika, że na podstawie referencji użytkownicy mogą ocenić wiarygodność artykułów Wikipedii (Lucassen i Schraagen, 2010).

Niektóre miary z tego wymiaru można podzielić na podgrupy, np. przy liczeniu referencji można brać pod uwagę każde odwołanie do dowolnego źródła lub liczyć tylko unikatowe referencje. Przedstawione w tej podsekcji miary mierzone są ilościowe, natomiast analiza jakościowa referencji przedstawia sekcja pt. „Miary jakości źródeł artykułów Wikipedii".

Do tego wymiaru należą następujące miary jakości: A20, A62, A88-A119.

5.6.9 Wielowymiarowe miary jakości

Z poprzednich podsekcji można zauważyć, że czasem określona miara może być powiązana z dwoma lub więcej wymiarami jakości. Na przykład liczba redaktorów może pokazywać obiektywność artykułu (inny punkt widzenia), ale dodatkowo może pomóc w mierzeniu relewantności treści (więcej użytkowników jest zainteresowanych tym tematem). Inny przykład to liczba obrazów. Z jednej strony obrazki mogą pomóc w ocenie obiektywności prezentowanego materiału, ale z drugiej strony możemy zmierzyć kompletność (ponieważ artykuły na określony temat powinny zawierać obrazy) i styl (na przykład, aby uniknąć pisania materiału w postaci tekstu, autorzy artykułu postanowili dodać więcej zdjęć). Liczba szablonów cytowania (Dang i Ignat, 2016a; Shen i in., 2017; Warncke-Wang i in., 2015; Zhang i in., 2018) może pomóc w mierzeniu liczby referencji (wiarygodność), a także w jakim stopniu informacje o źródle są dostępne dla czytnika (kompletność).

Tabela 5.7 pokazuje liczbę wspólnych miar związanych z różnymi wymiarami jakości artykułów Wikipedii. Największa liczba miar związana z relewancją. Największą liczbę wspólnych miar posiadają wymiary czytelność oraz kompletność. Miary relewancji dodatkowo mogą należeć prawie do wszystkich innych wymiarów jakości. Miary kompletności mają powiązania z czytelnością, obiektywnością, relewancją, stylem oraz wiarygodnością.

Tabela 5.7. Liczba wspólnych miar związanych z różnymi wymiarami jakości artykułów Wikipedii

	Akt.	**Czyt.**	**Komp.**	**Obj.**	**Popyt**	**Rel.**	**Styl**	**Wiar.**
Aktualność	**23**	0	0	16	0	16	0	0
Czytelność	0	**30**	24	0	0	2	17	1
Kompletność	0	24	**37**	8	0	2	16	3
Obiektywność	16	0	8	**32**	0	21	3	0
Popyt	0	0	0	0	**4**	4	0	0
Relewancja	16	2	2	21	4	**53**	0	1
Styl	0	17	16	3	0	0	**23**	3
Wiarygodność	0	1	3	0	0	1	3	**34**

Źródło: Opracowanie własne.

5.7 Podsumowanie

W tym rozdziale zostały omówione miary jakości artykułów Wikipedii w różnych językach. Dodatkowo zostały pokazane sposoby oraz źródła ekstrakcji tych miar. Miary zostały przypisane

do poszczególnych wymiarów jakości, takich jak: aktualność, czytelność, kompletność, obiektywność, relewantność, styl, wiarygodność. Niektóry miary mogą należeć do dwóch i więcej wymiarów jakości. Przedstawione miary będą używane do budowanie modeli jakości artykułów Wikipedii oraz infoboksów. Dodatkowo zostaną określone najważniejsze miary w modelach jakości różnych wersji językowych.

Rozdział 6

Budowanie modeli jakości artykułów Wikipedii

W niniejszym rozdziale zostały opisane zbudowane modele jakości artykułów Wikipedii w różnych wersjach językowych przy użyciu wszystkich miar jakości, które były opisane w poprzednim rozdziale. W celu wybrania metody z największą precyzją została przeprowadzona ewaluacja ponad 20 algorytmów klasyfikacji na różnych zbiorach danych.

Od tego rozdziału ocena jakości artykułu według różnego sposobu obliczania odznaczana jako J_x (lub Jx), gdzie x - numer (indeks) modelu, na podstawie którego został oceniony artykuł. W zależności od modelu jakość może być wartością dychotomiczną, nominalną lub ciągłą.

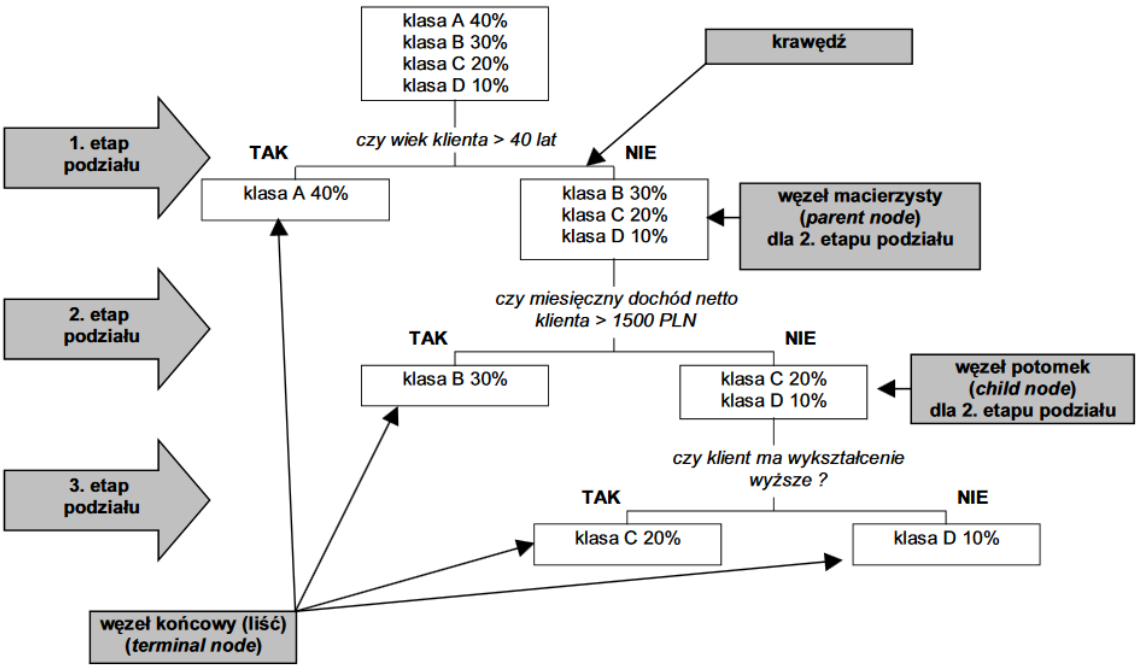
6.1 Wprowadzenie

Eksploracja danych (data mining) polega na wykorzystaniu określonych metod analitycznych do odkrywania ukrytych zależności, wzorców i reguł w ogromnych zasobach danych i informacji. Rozmaite metody i techniki w tym obszarze mogą pomóc w odnalezieniu najistotniejszych miar, które wpływają na ocenę artykułu w Wikipedii.

Wcześniejsze badania pokazały efektywność wykorzystywania drzew klasyfikacyjnych (decyzyjnych) w zadaniach automatycznej identyfikacji klasy jakości artykułów na podstawie ich atrybutów ilościowych (Lewoniewski i in., 2015, 2016; Warncke-wang i in., 2013; Y. Xu i Luo, 2011). W przypadku tych metod nie trzeba brać pod uwagę rozkładów danych – mogą one być nawet ze sobą skorelowane. Ponadto wygenerowane reguły logiczne są często łatwiejsze w interpretacji niż różnego rodzaju funkcje.

Powszechne stosowanie drzew decyzyjnych rozpoczęło się wraz z rozwojem algorytmu CART (Classification And Regression Tree) (Breiman, Friedman, Olshen i Stone, 1984). Jest to podstawowy algorytm, z wykorzystaniem którego może być zbudowane drzewo decyzyjne i jest podstawą wielu algorytmów budowy różnych drzew decyzyjnych.

Drzewo klasyfikacyjne buduje się w sposób rekurencyjny od korzenia do liści, a wewnętrzne węzły opisują sposób dokonania podziału na jednorodne klasy. Na rysunku 6.1 pokazana jest przykładowa budowa takiego drzewa. Algorytmy drzew klasyfikacyjnych różnią się m.in. kryteriami utworzenia kolejnego węzła lub liścia oraz testem dla węzła, który decyduje o złożoności drzewa decyzyjnego. Najbardziej popularne jest wykorzystanie przy tym entropii lub współczynnika Giniego do pomiaru nierównomierności rozkładu (Chistiakov, 2013). Przykładowe drzewo klasyfikacyjne pokazano na rysunku 6.1.



Rysunek 6.1. Przykład drzewa klasyfikacyjnego.

Źródło: (Łapczyński, 2003)

Istotnym postępem było wprowadzenie entropii do pomiaru różnorodności w drzewach. Pierwszym wykorzystującym entropię algorytmem do budowy drzew decyzyjnych był ID3 Quinlana. Jednak posiadał on pewne wady: niepotrzebny rozrost drzewa i brak mechanizmów przeciwdziałających zjawisku nadmiernego dopasowania (*overfitting*), a w konsekwencji wysoki poziom błędów dla zbiorów oceniających. Z czasem powstał algorytm C4.5, który jest roz-

szerzeniem ID3 i zapobiega problemom oryginalnej metody (Quinlan, 1993). Później została przedstawiona kolejna wersja algorytmu - C5.0 (Freund i Mason, 1999).

Szczególną uwagę należy zwrócić na algorytm lasu losowego (Random Forest), który nadaje się do rozwiązywania nie tylko zadań klasyfikacyjnych (Breiman, 2001). W metodzie tej budowane są stosunkowo proste drzewa decyzyjne, a rezultat całego modelu wyznaczany jest przez głosowanie lub uśrednianie.

6.2 Dobór algorytmów eksploracji danych oraz zbioru danych

Biorąc pod uwagę wcześniejsze badania w tym obszarze, do budowy modeli prognozowania jakości artykułów Wikipedii zostało wybranych ponad 20 algorytmów klasyfikacyjnych. Lista i opis większości modeli przedstawiona jest w tabeli 6.1. Wszystkie algorytmy działały na domyślnych ustawieniach w programie WEKA.

W celu ekstrakcji miar zostały wykorzystane pliki kopii zapasowej Wikipedii każdej z rozpatrywanych wersji językowych z danymi na lipiec 2018 roku. W szczególności zostały pobrane pliki opisane w sekcji 5.2 „Metody oraz źródła ekstrakcji miar". Dodatkowo zostały pobrane dane związane z pomiarem liczby odwiedzających każdej strony Wikipedii.

Zostały przygotowane własne algorytmy w języku Python, które na podstawie przetworzenia pobranych plików do każdego z rozpatrywanych artykułów przypisywały określone miary. Większa część z tych miar dotyczy artykułów oraz infoboksów. Miary te były później wykorzystywane jako zmienne niezależne w modelach.

6.3 Dobór zmiennej zależnej

Na podstawie ocenionych artykułów Wikipedii można zbudować predykcyjny model jakości. W tym przypadku model musi na podstawie znajomości wartości rożnych miar nauczyć się określać przynależność nieocenionych artykułów do określonych klas jakości. Wówczas w takim modelu zmienna zależna będzie nominalna.

Artykuły można również dobierać w grupy klas. Wtedy liczba kategorii będzie redukowana. Na potrzeby budowy modeli będą używane dwie grupy klas jakości: „Kompletne" oraz „Niekompletne". W takim modelu zmienna zależna będzie dychotomiczna.

Tabela 6.1. Modele klasyfikacyjne wykorzystane w analizie

Nazwa	Opis
Random Forest	Tworzenie wielu drzew decyzyjnych. Zbudowanie konsylium ekspertów z losowych drzew decyzyjnych (Breiman, 2001)
Logistic Model Trees	Drzewa klasyfikacyjne z funkcjami regresji logistycznej na liściach (Landwehr, Hall i Frank, 2005; Sumner, Frank i Hall, 2005)
C4.5	Następca ID3, w którym dla budowy drzewa wybiera się zmienną, która maksymalizuje redukcję entropii (Quinlan, 1993)
C4.5 graft	Budowanie przedłużonych drzew klasyfikacyjnych C4.5 (Webb, 1999)
C5.0	Następca algorytmu C4.5 (Freund i Mason, 1999)
FT	Drzewa funkcjonalne (Gama, 2004; Landwehr, Hall i Frank, 2005)
BFTree	Budowanie klasyfikatora najlepszego drzewa decyzyjnego. Używa podziału binarnego dla atrybutów nominalnych oraz liczbowych (Friedman, Hastie i Tibshirani, 2000; Shi, 2007)
CART	Classification And Regression Tree - budowa binarnych drzew decyzyjnych przy wykorzystaniu wskaźnika Gini do wyboru zmiennych do punktów podziału. (Breiman, Friedman, Olshen i Stone, 1984)
REPTree	Przycinanie drzewa opiera się na metodzie redukcji błędów
LADTree	Przemienne drzewo decyzyjne stosujące strategię LogitBoost (Holmes, Pfahringer, Kirkby, Frank i Hall, 2001)
ADTree	Przemienne drzewo decyzyjne (Freund i Mason, 1999)
NBTree	Drzewa decyzyjne z klasyfikatorem bayesowskim na liściach (Kohavi, 1996)
DecisionTable	Uproszczony klasyfikator do tablic decyzyjnych. (Kohavi, 1995)
Random Tree	Konstruuje wiele drzew decyzyjnych losowo. Tworzenie węzła bez jakiejkolwiek kontroli równomierności rozkładu
AdaBoostM1	Wzmocnione drzewa decyzyjne (Freund i Schapire, 1996)
SMO	Sequential Minimal Optimization (SMO) - wersja SVM dla zadań klasyfikicyjnych (Hastie i Tibshirani, 1998; Keerthi, Shevade, Bhattacharyya i Murthy, 2001; Platt, 1998)
DecisionStump	Jednopoziomowe drzewa decyzyjne (składające się tylko z korzenia i liści)
MLP	MultilayerPerceptron - algorytm sieci neuronowych. Wykorzystuje propagację wsteczną do klasyfikowania przypadków.

6.3.1 Nominalna zmienna zależna

Najpierw zostały zbudowane modele jakości, gdzie zmienną zależną była przynależność artykułów do określonych przez użytkowników Wikipedii klas jakości. W zależności od wersji językowej Wikipedii liczba takich klas w zbiorze danych może być różna. Np. dla angielskiej Wikipedii jest 6 wartości zmiennej zależnej: FA, GA, B-class, C-class, Start, Stub. Artykuły z klasy A zazwyczaj posiadają dodatkowo wyższą ocenę FA bądź GA, dlatego ta klasa nie będzie rozpatrywana oddzielnie, podobnie jak to było robione w innych badaniach (Dang i Ignat, 2016b; Halfaker, 2017; Shen i in., 2017).

W celu zbudowania zbilansowanej próby, liczba artykułów z poszczególnych klas była dobrana biorąc pod uwagę liczebność najmniejszej klasy jakości - FA, która zawierała 5344 artykułów. W wyniku losowego doboru 5 000 artykułów z każdej klasy otrzymano zbiór 30 000 artykułów.

Niektóre algorytmy wymagają konwersji wartości kategorialnej zmiennej zależnej na liczby (np. algorytmy klasyfikacji w bibliotece scikit-learn w języku programowania Python). W takim przypadku ważne jest, aby zachować kolejność klas zgodnie z malejąca lub rosnącą jakością (Dalip i in., 2017).

6.3.2 Dychotomiczna zmienna zależna

Podobnie jak w innych badaniach do budowania modelu może być stosowana dychotomiczna zmienna objaśniana (Lewoniewski i in., 2015; Lex i in., 2012; Su i Liu, 2015; Warncke-wang i in., 2013; Y. Xu i Luo, 2011) i jakość może być modelowana jako prawdopodobieństwo przynależności do jednej z dwóch klas:

- `Kompletne` artykuły: klasy FA i GA,
- `Niekompletne` artykuły: wszystkie inne – rozwijające się (które należy dopracować) oraz nieocenione artykuły.

W niektórych pracach można spotkać podział uwzględniający tylko artykuły klasy FA (jako wzorowych) i innych losowo dobranych artykułów (Su i Liu, 2015).

Dobór próby uczącej ponownie odbywał się z uwzględnieniem liczby artykułów w najmniejszej klasie. Na przykład w wersji angielskiej najmniejszą liczbę artykułów zawierała najwyższa klasa FA - 5 344 artykułów. W związku z tym wybrano losowo 5 000 artykułów z klasy FA oraz 5 000 artykułów z klasy GA do umieszczenia w kategorii „Kompletne”, co dało łącznie 10 000

artykułów. Taką samą liczbę artykułów wylosowano do kategorii „Niekompletne”, przy czym w celu zachowania równoważności tej grupy, z pozostałych 4 klas jakości (B, C, Start, Stub) wylosowano odpowiednio 2 500 artykułów. Cała próba ucząca wykorzystana do budowania modeli jakości przy użyciu dychotomicznej zmiennej zależnej wyniosła 20 000 artykułów.

6.4 Ewaluacja algorytmów klasyfikacyjnych

Wśród rozpatrywanych wersji językowych największa liczbę ocenionych artykułów zawiera angielska Wikipedia. Dodatkowym interesującym przypadkiem jest rosyjskojęzyczna wersja Wikipedii, która jest drugą największa wersją z rozpatrywanych w tej pracy. Co więcej, posiada również rozwinięty system oceny jakości artykułów oraz większą liczbę ocenionych artykułów w porównaniu do polskiej, ukraińskiej oraz białoruskiej. Należy jednak brać również pod uwagę różnice w systemach ocen w angielskiej oraz rosyjskiej Wikipedii - różna liczba klas jakości, inne nazwy ocen oraz zasady ich przyznawania.

W związku z tym w kolejnych podsekcjach będą oddzielnie opisane wyniki ewaluacji algorytmów klasyfikacyjnych na podstawie zbiorów danych z angielskiej oraz rosyjskiej Wikipedii. Przy ewaluacji były wykorzystane algorytmy zaimplementowane w pakiecie statystycznym WEKA (Hall i in., 2009; Holmes, Donkin i Witten, 1994), z wykorzystaniem domyślnych ustawień oraz 10-krotnej walidacji krzyżowej.

Należy zaznaczyć, że w zależności od typu zmiennej zależnej, niektóre algorytmy nie były stosowane: DecisionStump, MultilayerPerceptron, AdaBoostM1 były używane tylko w modelach z dychotomiczną zmienną zależną. To wynika z zasady działania tych algorytmów i koniecznością dopasowania zmiennej zależnej do nich.

6.4.1 Angielska Wikipedia

W pierwszej kolejności zostały zbudowane modele wykorzystujące zbiór uczący z dwiema klasami jakości:

- Kompletne - 5000 artykułów z klasy FA oraz 5000 artykułów z klasy GA. Razem - 10000 art.
- Niekompletne - 2500 artykułów z każdej z pozostałych czterech klas jakości: B, C, Start, Stub. Razem - 10000 art.

Dla każdego artykułu zostały wyekstrahowane miary, które były opisane w poprzednim rozdziale („Miary oraz wymiary jakości artykułów Wikipedii"). Na podstawie tych danych zostały zbudowane modele z wykorzystaniem różnych wcześniej opisanych algorytmów (patrz tabelę 6.1)

Do oceny jakości klasyfikatorów wykorzystywane mogą być różne narzędzia. Jedno z podstawowych - macierz błędu, która może być stosowana dla pokazania rozbieżności pomiędzy klasami, do których należą artykuły (klasy rzeczywiste) oraz klasami, które były określone przez model (wynik). Innymi słowy, ta macierz może wykryć, ile z oryginalnie oznaczonych jako „Kompletne" („NieKompletne") zostanie omyłkowo zaklasyfikowana jako „Niekompletne" („Kompletne"). Tabela 6.2 pokazuje macierz błędów (tablica pomyłek) tego modelu oceny jakości przy wykorzystaniu algorytmu lasu losowego (Random Forest).

Tabela 6.2. Macierz błędów w modelu predykcji jakości w angielskiej Wikipedii przy użyciu dychotomicznej zmiennej zależnej z wykorzystaniem algorytmu RandomForest.

		Wynik modelu	
		Kompletne	**NieKompletne**
Klasa rzeczywista	**Kompletne**	9615	385
	Niekompletne	804	9196

Źródło: Obliczenia własne przy użyciu pakiety statystycznego WEKA.

Na podstawie macierzy błędów można obliczyć różne wskaźniki jakości modelu. Niektóre z tych wskaźników zostały opisane w tabeli 6.3.

W tabeli 6.4 zostały pokazane średnie ważone wskaźniki jakości poszczególnych modeli na zbiorze danych z angielskiej Wikipedii z dychotomiczną zmienną zależną.

Wyniki pokazują, że w zależności od algorytmu klasyfikacyjnego, można osiągnąć ponad 90-procentową precyzję. Najlepsze wskaźniki wykazał algorytm lasu losowego.

Tabela 6.5 pokazuje szczegółową informację na temat wskaźników jakości w modelu predykcji jakości w angielskiej Wikipedii przy użyciu dychotomicznej zmiennej zależnej z wykorzystaniem algorytmu RandomForest.

Następnie zostały zbudowane modele jakości, w których dla każdej klasy jakości przypisano oddzielna kategoria. W angielskiej Wikipedii takich grup artykułów jest 6: FA, GA, B, C, Start, Stub. Dla każdej z tych klas zostało losowo dobrane po 5 000 artykułów. Razem zbiór danych liczył 30 000 artykułów. Dla każdego z wybranych artykułów zostały wyekstrahowane ponad 100 różnych miar jakości, opisanych w rozdziale „Miary oraz wymiary jakości artykułów Wikipedii".

Tabela 6.3. Wskaźniki jakości modelu klasyfikacyjnego.

Wskaźnik	Opis
TP (True Positive)	Prawdziwie pozytywna. W rozpatrywanym przykładzie to jest wartość 9615, która pokazuje ile artykułów oryginalnie oznaczonych jako „Kompletne" zostały zaklasyfikowany jako „Kompletne" przez model. TP Rate oznacza stosunek artykułów oznaczonych przez model jako „Kompletne" do liczby wszystkich artykułów, które rzeczywiście do tej klasy należą.
TN (True Negative)	Prawdziwie negatywna. W rozpatrywanym przykładzie to jest wartość 9196, która pokazuje ile artykułów oryginalnie oznaczonych jako „Niempletne" zostały zaklasyfikowany jako „NieKompletne".
FP (False Positive)	Fałszywie pozytywna czy błąd pierwszego rodzaju. W rozpatrywanym przykładzie to jest wartość 804, która pokazuje ile artykułów oryginalnie oznaczonych jako „NieKompletne" zostały omyłkowo zaklasyfikowany jako „Kompletne". FP Rate oznacza stosunek artykułów omyłkowo oznaczonych przez model jako „Kompletne" do liczby wszystkich artykułów, które należą do klasy „NieKompletne".
FN (False Negative)	Fałszywie negatywna czy błąd drugiego rodzaju. W rozpatrywanym przykładzie to jest wartość 385, która pokazuje ile artykułów oryginalnie oznaczonych jako „Kompletne" zostały zaklasyfikowany omyłkowo jako „NieKompletne" przez model.
Precision	Precyzja modelu, liczona na podstawie wzoru: $Precision = TP/(TP + FP)$
Recall	Czułość modelu, liczona na podstawie wzoru: $Recall = TP/(TP + FN)$
F-measure	Miara liczona na podstawie wzoru: $\frac{2*Precision*Recall}{Precision+Recall}$
MCC	współczynnik korelacji Matthews liczony na podstawie wzoru: $MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
ROC (Receiver Operating Characteristics)	Prawdopodobieństwo, że badany model predykcyjny oceni wyżej losowy element klasy pozytywnej („Kompletne") od losowego elementu klasy negatywnej („NieKompletne"). ROC - to funkcja punktu odcięcia, przedstawia zmienność TP Rate w zależności od FP Rate.
PRC (Precision-Recall Curve)	Pokazuje zależność między precyzją (Precision) a czułością (Recall) dla każdego możliwego odcięcia na wykresie, gdzie na osi OX pokazana precyzja oraz na osi OY - czułość modelu. W odróżnieniu od ROC, ta wartość może być bardziej przydatna, jeżeli badane jest zachowanie klasyfikatora tylko w ramach określonej klasy (Saito i Rehmsmeier, 2015).

Tabela 6.4. Wskaźniki jakości modeli klasyfikacyjnych na zbiorze artykułów z angielskiej Wikipedii przy użyciu dychotomicznej zmiennej zależnej. Sortowano według precyzji.

Algorytm	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
RandomForest	0,941	0,941	0,941	0,882	0,985	0,984
Bagging	0,934	0,933	0,933	0,868	0,980	0,979
RandomSubSpace	0,934	0,933	0,933	0,866	0,979	0,979
LMT	0,932	0,932	0,931	0,863	0,975	0,968
RandomCommittee	0,930	0,930	0,930	0,860	0,978	0,973
ClassificationViaRegression	0,924	0,923	0,923	0,847	0,972	0,969
PART	0,923	0,922	0,922	0,845	0,927	0,905
REPTree	0,914	0,913	0,913	0,827	0,956	0,946
AttributeSelectedClassifier	0,911	0,910	0,910	0,821	0,916	0,889
J48	0,910	0,910	0,910	0,820	0,896	0,864
MultiClassClassifier	0,908	0,908	0,907	0,815	0,963	0,958
IterativeClassifierOptimizer	0,903	0,901	0,901	0,804	0,962	0,960
LogitBoost	0,903	0,901	0,901	0,804	0,962	0,960
SimpleLogistic	0,901	0,901	0,901	0,802	0,963	0,959
SMO	0,900	0,899	0,899	0,799	0,899	0,859
FilteredClassifier	0,891	0,890	0,890	0,780	0,908	0,878
DecisionTable	0,886	0,884	0,884	0,770	0,946	0,942
AdaBoostM1	0,877	0,877	0,877	0,754	0,950	0,942
RandomTree	0,877	0,877	0,876	0,753	0,877	0,830
MultilayerPerceptron	0,865	0,851	0,850	0,716	0,949	0,946
DecisionStump	0,853	0,848	0,848	0,702	0,848	0,797
OneR	0,844	0,843	0,842	0,687	0,843	0,789
HoeffdingTree	0,827	0,809	0,806	0,636	0,793	0,767
BayesNet	0,825	0,823	0,823	0,648	0,882	0,855
NaiveBayes	0,710	0,659	0,636	0,365	0,798	0,737
RandomizableFilteredClassifier	0,643	0,643	0,643	0,286	0,643	0,595

Źródło: Obliczenia własne w programie WEKA.

Tabela 6.5. Wskaźniki jakości w modelu predykcji jakości w angielskiej Wikipedii przy użyciu dychotomicznej zmiennej zależnej z wykorzystaniem algorytmu RandomForest.

Klasa	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Kompletne	0,962	0,08	0,923	0,962	0,942	0,882	0,985	0,982
NieKompletne	0,92	0,039	0,96	0,92	0,939	0,882	0,985	0,986
Średnia ważona	**0,941**	**0,059**	**0,941**	**0,941**	**0,941**	**0,882**	**0,985**	**0,984**

Źródło: Obliczenia własne przy użyciu pakiety statystycznego WEKA.

Podobnie jak w przypadku klasyfikacji binarnej, dla większej niż 2 liczby klas można zbudować macierz błędów, która pokazuje rozbieżności pomiędzy rzeczywistymi klasami oraz klasami, określonymi przez model. Tabela 6.6 przedstawia taką macierz dla 6 klas jakości w angielskiej Wikipedii przy użyciu algorytmu lasa losowego. Warto zwrócić uwagę, że największe rozbieżności pomiędzy liczbami artykułów rzeczywistych oraz oznaczonych przez model występują w sąsiednich według jakości klasach. To może oznaczać brak sztywnych granic pomiędzy kryteriami jakości w bliskich klasach. Najniższa predykcja modelu wykazana przy wyznaczeniu jakości artykułów z pośrednich klas: B oraz C.

Tabela 6.6. Macierz błędów w modelu predykcji jakości w angielskiej Wikipedii przy użyciu nominalnej zmiennej zależnej z wykorzystaniem algorytmu RandomForest.

		Wynik modelu					
		FA	**GA**	**B**	**C**	**Start**	**Stub**
Klasa rzeczywista	**FA**	**4648**	284	60	8	0	0
	GA	970	**3657**	226	119	27	1
	B	295	657	**2141**	1244	610	53
	C	71	348	1206	**2185**	1115	75
	Start	12	77	290	974	**2917**	730
	Stub	0	1	20	107	991	**3881**

Źródło: Obliczenia własne przy użyciu pakiety statystycznego WEKA.

W tym przypadku macierz błędów również może być stosowana do obliczeń miar jakości modelu. Do obliczenia wskaźników (np. ROC), które zazwyczaj są stosowane do oceny algorytmów klasyfikacji binarnej, wykorzystana jest średnia ważona, która uwzględnia obliczenia tych wskaźników z punktu widzenia poszczególnych klas. Np. przy obliczeniu ROC dla klasy FA, wszystkie artykuły są dzielone na dwie grupy: FA oraz inne klasy (jako jedna wspólna).

W tabeli 6.7 zostały pokazane wskaźniki jakości poszczególnych algorytmów na zbiorze danych z angielskiej Wikipedii z nominalną zmienną zależną. Do obliczenia każdego wskaźnika została użyta średnia ważona wartości wskaźników dla każdej z rozpatrywanych klas jakości artykułów.

Wyniki analizy wskaźników jakości modelu pokazują przewagę algorytmu Random Forest. Ten algorytm posiada największe wartości wszystkich rozpatrywanych wskaźników.

Warto zaznaczyć że wykorzystanie dodatkowych miar jakości, zaproponowanych w niniejszej rozprawie, pozwala na zbudowanie bardziej precyzyjnych modeli, niż w innych badaniach

Tabela 6.7. Wskaźniki modeli klasyfikacyjnych na zbiorze artykułów z angielskiej Wikipedii przy użyci nominalnej zmiennej zależnej. Sortowano według precyzji.

Algorytm	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
RandomForest	0,642	0,648	0,642	0,574	0,913	0,688
LMT	0,618	0,622	0,618	0,544	0,900	0,641
RandomSubSpace	0,613	0,622	0,614	0,542	0,903	0,663
Bagging	0,611	0,619	0,613	0,538	0,899	0,660
RandomCommittee	0,601	0,605	0,601	0,524	0,879	0,612
ClassificationViaRegression	0,595	0,604	0,599	0,521	0,883	0,626
JRip	0,583	0,527	0,512	0,452	0,815	0,492
SimpleLogistic	0,581	0,593	0,583	0,505	0,885	0,601
IterativeClassifierOptimizer	0,571	0,584	0,571	0,493	0,884	0,613
LogitBoost	0,571	0,584	0,571	0,493	0,884	0,613
REPTree	0,570	0,578	0,572	0,489	0,858	0,568
PART	0,567	0,567	0,567	0,480	0,759	0,455
SMO	0,563	0,573	0,563	0,481	0,846	0,484
MultiClassClassifier	0,562	0,577	0,564	0,484	0,873	0,575
J48	0,539	0,539	0,539	0,447	0,740	0,423
AttributeSelectedClassifier	0,533	0,535	0,534	0,441	0,739	0,422
FilteredClassifier	0,505	0,516	0,508	0,413	0,757	0,431
DecisionTable	0,504	0,518	0,504	0,414	0,841	0,509
RandomTree	0,501	0,501	0,501	0,401	0,700	0,358
BayesNet	0,485	0,497	0,471	0,386	0,822	0,487
OneR	0,382	0,398	0,385	0,269	0,639	0,288
HoeffdingTree	0,336	0,351	0,249	0,198	0,631	0,281
NaiveBayes	0,316	0,310	0,264	0,165	0,718	0,316
RandomizableFilteredClassifier	0,288	0,289	0,289	0,147	0,573	0,213

Źródło: Obliczenia własne w programie WEKA.

w tym obszarze (Dang i Ignat, 2016b; Halfaker, 2017; Shen i in., 2017; Warncke-wang i in., 2013; Warncke-Wang i in., 2015).

Tabela 6.8. Wskaźniki jakości w modelu predykcji jakości w angielskiej Wikipedii przy użyciu nominalnej zmiennej zależnej z wykorzystaniem algorytmu RandomForest.

Klasa	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
FA	0,93	0,054	0,775	0,93	0,845	0,816	0,984	0,916
GA	0,731	0,055	0,728	0,731	0,73	0,675	0,95	0,818
B	0,428	0,072	0,543	0,428	0,479	0,393	0,852	0,506
C	0,437	0,098	0,471	0,437	0,453	0,349	0,845	0,459
Start	0,583	0,11	0,515	0,583	0,547	0,451	0,876	0,547
Stub	0,776	0,034	0,819	0,776	0,797	0,758	0,969	0,885
Średnia ważona	**0,648**	**0,07**	**0,642**	**0,648**	**0,642**	**0,574**	**0,913**	**0,688**

Źródło: Obliczenia własne przy użyciu pakiety statystycznego WEKA.

6.4.2 Rosyjska Wikipedia

Podobnie jak i w przypadku angielskiej Wikipedii, najpierw zostały zbudowane modele jakości dla dwóch grup artykułów. W związku z tym, że w wersji rosyjskiej istnieją inne standardy jakości oraz inna liczba artykułów w najmniejszej klasie, zawartość zbioru danych wygląda następująco:

- Kompletne - 600 artykułów z każdej z 3 najwyższych klas jakości: FA, GA, SA. Razem - 1800 art.
- Niekompletne - 450 artykułów z każdej z 4 pozostałych klas: I, II, III, IV. Razem - 1800 art.

W tabeli 6.9 zostały pokazane wskaźniki jakości poszczególnych algorytmów na zbiorze danych z rosyjskiej Wikipedii przy wykorzystaniu dychotomicznej zmiennej zależnej.

Tabela 6.9. Wskaźniki modeli klasyfikacyjnych na zbiorze artykułów z angielskiej Wikipedii przy użyci dychotomicznej zmiennej zależnej. Sortowano według precyzji.

Algorytm	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
RandomForest	0,932	0,930	0,930	0,862	0,981	0,981
Bagging	0,929	0,927	0,927	0,856	0,978	0,977
RandomCommittee	0,928	0,928	0,928	0,856	0,975	0,968
RandomSubSpace	0,925	0,922	0,922	0,847	0,976	0,976
LMT	0,919	0,918	0,918	0,837	0,959	0,948
ClassificationViaRegression	0,916	0,916	0,915	0,832	0,966	0,962
PART	0,916	0,916	0,916	0,832	0,913	0,887
REPTree	0,912	0,910	0,910	0,823	0,954	0,943
AttributeSelectedClassifier	0,904	0,903	0,903	0,807	0,910	0,885
J48	0,904	0,904	0,904	0,808	0,897	0,869
FilteredClassifier	0,904	0,903	0,903	0,807	0,919	0,898
MultiClassClassifier	0,904	0,904	0,903	0,807	0,955	0,948
IterativeClassifierOptimizer	0,901	0,897	0,897	0,798	0,956	0,952
LogitBoost	0,901	0,897	0,897	0,798	0,956	0,952
AdaBoostM1	0,890	0,886	0,886	0,777	0,954	0,950
SMO	0,886	0,885	0,885	0,770	0,885	0,841
SimpleLogistic	0,882	0,882	0,882	0,764	0,944	0,941
DecisionTable	0,882	0,878	0,878	0,760	0,940	0,935
RandomTree	0,873	0,873	0,873	0,746	0,873	0,826
MultilayerPerceptron	0,871	0,857	0,855	0,728	0,933	0,926
DecisionStump	0,862	0,839	0,837	0,701	0,839	0,793
OneR	0,851	0,843	0,842	0,693	0,843	0,792
HoeffdingTree	0,828	0,822	0,821	0,650	0,842	0,809
BayesNet	0,819	0,818	0,818	0,636	0,891	0,870
NaiveBayes	0,734	0,694	0,681	0,427	0,806	0,751
RandomizableFilteredClassifier	0,662	0,661	0,661	0,323	0,661	0,611

Źródło: obliczenia własne w programie WEKA

Wyniki pokazują, że przy wykorzystaniu dychotomicznej zmiennej zależnej najwyższe wskaźniki jakości modelu wykazał algorytm lasu losowego. Tabela 6.10 przedstawia macierz błędów, otrzymana przy użyciu danego algorytmu.

Tabela 6.10. Macierz błędów w modelu predykcji jakości w rosyjskojęzycznej Wikipedii przy użyciu dychotomicznej zmiennej zależnej z wykorzystaniem algorytmu RandomForest.

		Wynik modelu	
		Kompletne	**NieKompletne**
Klasa rzeczywista	**Kompletne**	2888	112
	Niekompletne	306	2694

Źródło: Obliczenia własne przy użyciu pakiety statystycznego WEKA.

Szczegółowe informacje na temat wskaźników jakości algorytmu lasu losowego na artykułach rosyjskojęzycznej Wikipedii przy użyciu dychotomicznej zmiennej zależnej przedstawia tabela 6.11.

Tabela 6.11. Wskaźniki jakości w modelu predykcji jakości w rosyjskojęzycznej Wikipedii przy użyciu dychotomicznej zmiennej zależnej z wykorzystaniem algorytmu RandomForest.

Klasa	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Kompletne	0,963	0,102	0,904	0,963	0,933	0,862	0,981	0,978
NieKompletne	0,898	0,037	0,96	0,898	0,928	0,862	0,981	0,983
Średnia ważona	**0,93**	**0,07**	**0,932**	**0,93**	**0,93**	**0,862**	**0,981**	**0,981**

Źródło: Obliczenia własne przy użyciu pakiety statystycznego WEKA.

Następnie zostały zbudowane modele przy użyciu nominalnej zmiennej zależnej. W tym przypadku, do każdej klasy zostały dobrano 600 artykułów.

W tabeli 6.12 zostały pokazane wskaźniki jakości poszczególnych algorytmów na zbiorze danych z rosyjskiej Wikipedii przy wykorzystaniu nominalnej zmiennej zależnej.

Algorytm lasu losowego wykazał w tym przypadku również najlepsze wskazniki. Macierz błędów pokazana jest w tabeli 6.13. Można zauważyć, że największe trudności dla modelu sprawia predykcja artykułów z pośrednich klas jakości: I, II oraz III.

Tabela 6.14 przedstawia wskaźniki jakości w modelu predykcji jakości w rosyjskojęzycznej Wikipedii przy użyciu dychotomicznej zmiennej zależnej z wykorzystaniem algorytmu RandomForest.

Tabela 6.12. Wskaźniki modeli klasyfikacyjnych na zbiorze artykułów z rosyjskiej Wikipedii przy użyci kategorialnej zmiennej zależnej. Sortowano według precyzji.

Algorytm	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
RandomForest	0,655	0,660	0,654	0,600	0,929	0,720
Bagging	0,641	0,644	0,641	0,583	0,921	0,700
RandomSubSpace	0,639	0,646	0,639	0,583	0,923	0,700
ClassificationViaRegression	0,628	0,632	0,629	0,569	0,910	0,672
RandomCommittee	0,627	0,630	0,626	0,566	0,901	0,652
JRip	0,616	0,573	0,562	0,515	0,858	0,538
IterativeClassifierOptimizer	0,616	0,623	0,617	0,556	0,913	0,662
LogitBoost	0,616	0,623	0,617	0,556	0,913	0,662
LMT	0,615	0,620	0,616	0,554	0,901	0,624
REPTree	0,607	0,609	0,606	0,542	0,883	0,599
FilteredClassifier	0,598	0,604	0,600	0,535	0,805	0,501
BayesNet	0,591	0,605	0,586	0,529	0,891	0,630
J48	0,588	0,589	0,588	0,520	0,772	0,469
PART	0,585	0,582	0,583	0,514	0,773	0,470
AttributeSelectedClassifier	0,573	0,573	0,572	0,501	0,774	0,469
DecisionTable	0,569	0,572	0,564	0,498	0,883	0,571
SimpleLogistic	0,569	0,583	0,569	0,505	0,896	0,607
MultiClassClassifier	0,568	0,583	0,570	0,505	0,893	0,611
SMO	0,555	0,569	0,555	0,488	0,859	0,479
RandomTree	0,533	0,531	0,531	0,453	0,726	0,376
HoeffdingTree	0,407	0,400	0,358	0,295	0,773	0,369
NaiveBayes	0,406	0,403	0,360	0,297	0,780	0,386
OneR	0,405	0,434	0,409	0,324	0,670	0,290
RandomizableFilteredClassifier	0,334	0,333	0,333	0,222	0,611	0,225

Źródło: Obliczenia własne w programie WEKA.

Tabela 6.13. Macierz błędów w modelu predykcji jakości w rosyjskojęzycznej Wikipedii przy użyciu kategorialnej zmiennej zależnej z wykorzystaniem algorytmu RandomForest.

		Wynik Modelu						
		FA	**GA**	**SA**	**I**	**II**	**III**	**IV**
Klasa rzeczywista	**FA**	**869**	116	0	15	0	0	0
	GA	224	**709**	35	30	2	0	0
	SA	1	56	**915**	12	14	2	0
	I	85	90	41	**455**	246	52	31
	II	16	37	42	214	**480**	178	33
	III	1	7	23	48	241	**482**	198
	IV	0	1	6	18	65	202	**708**

Źródło: Obliczenia własne przy użyciu pakiety statystycznego WEKA.

Tabela 6.14. Wskaźniki jakości w modelu predykcji jakości w rosyjskojęzycznej Wikipedii przy użyciu dychotomicznej zmiennej zależnej z wykorzystaniem algorytmu RandomForest.

Klasa	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
FA	0,869	0,055	0,727	0,869	0,791	0,757	0,979	0,872
GA	0,709	0,051	0,698	0,709	0,703	0,654	0,957	0,806
SA	0,915	0,025	0,862	0,915	0,887	0,869	0,992	0,963
I	0,455	0,056	0,574	0,455	0,508	0,441	0,877	0,558
II	0,48	0,095	0,458	0,48	0,469	0,378	0,863	0,448
III	0,482	0,072	0,526	0,482	0,503	0,425	0,886	0,574
IV	0,708	0,044	0,73	0,708	0,719	0,673	0,949	0,794
Średnia ważona	**0,66**	**0,057**	**0,654**	**0,66**	**0,654**	**0,599**	**0,929**	**0,716**

Źródło: Obliczenia własne przy użyciu pakiety statystycznego WEKA.

6.4.3 Wnioski z ewaluacji algorytmów

Przeprowadzone analizy pokazały przewagę algorytmu lasu losowego (Random Forest) w predykcji jakości artykułów w różnych wersjach językowych Wikipedii przy dychotomicznej oraz kategorialnej zmiennej zależnej. Główną wadą algorytmu w porównaniu do innych drzew decyzyjnych jest brak wizualnej reprezentacji procesu decyzyjnego i złożoność interpretacji decyzji.

Niemniej jednak algorytm oblicza miarę ważności zmiennych objaśniających i macierzy podobieństwa obserwacji, co może zrekompensować ten niedobór. Ważność zmiennych objaśniających pozwala wyróżnić najbardziej znaczące cechy artykułów, który wpływają na ich jakość.

Taka analiza w naszych badaniach pokazała różnice w modelach jakości artykułów w różnych wersjach językowych Wikipedii. Macierz podobieństwa RandomForest umożliwia zastosowanie metod, które nie są bezpośrednio związane z klasyfikacją i regresją - takich jak skalowanie wielowymiarowe, analiza skupień, identyfikacja anomalnych obserwacji i inne. Te możliwości lasu losowego w dużej mierze przyczyniły się do wzrostu popularności tego algorytmu.

Do najważniejszych zalet algorytmu Random Forest można zaliczyć:

- możliwość wykorzystania dużej liczby zmiennych niezależnych (miar), które mogą być ze sobą skorelowane (Breiman, 2001),
- odporność na braki danych oraz wartości odstające,
- możliwość badania interakcji między zmiennymi niezależnymi,
- stosunkowa łatwość interpretacji wyników,
- stabilność oraz odporność na przeuczenie (Hastie, Tibshirani i Friedman, 2009).

6.5 Ważność miar w modelach jakości

Algorytm Random Forest posiada możliwość obliczenia ważności miar, które zostały użyte do budowania modeli jakości. Ważność może być obliczana na różny sposób. Na przykład, pakiet statystyczny WEKA pozwala obliczać ważność na podstawie następujących wskaźników (oddzielnie):

- miara Gini'ego - średnie zmniejszenia niejednorodności zbioru przy użyciu wybranej miary,
- liczby węzłów używających wybranej miary przy budowaniu drzew klasyfikacyjnych.

Wskaźnik Gini'ego niespójności węzła jest często używany w sytuacjach, gdy zmienna zależna jest zmienną nominalną lub dychotomiczną (jako szczególny przypadek zmiennej nominalnej). W celu obliczenia tego wskaźnika dla zbioru artykułów z K kategoriami, niech $i \in \{1, 2, ..., K\}$, i niech p_i będzie stosunek liczby artykułów oznaczonych kategorią i do liczby wszystkich artykułów w naszym zbiorze. Wtedy wskaźnik Gini'ego dla miary m obliczany według wzoru (Venables i Ripley, 2002):

$Gini(m) = 1 - \sum_{i=1}^{K} {p_i}^2$

W związku z tym, że algorytm lasu losowego tworzy wiele drzew klasyfikacyjnych, dla każdego z tych drzew konkretna miara może mieć różny wskaźnik Gini'ego. W związku z tym, pakiet statystyczny WEKA oblicza średnią wartość tego wskaźnika.

Innym sposobem obliczania ważności miary jest uwzględnienie liczby węzłów drzew klasyfikacyjnych, które tę miarę używają. Każde drzewo w algorytmie lasu losowego budowane jest na innym, losowo wybranym podzbiorze zbioru danych uczących, a wybierany jest najlepszy podział z losowego podzbioru predyktorów (miar).

Dla obliczenia ważności W miary jakości m w tej pracy został użyty wskaźnik, będący iloczynem średniego wskaźnika Gini'ego dla tej miary oraz liczby węzłów drzew klasyfikacyjnych używających tę miarę $L(m)$:

$W(m) = Gini(m) * L(m)$

W celu bardziej wygodnego porównywania ważności miar w różnych modelach w niniejszej pracy wartości $W(m)$ będą znormalizowane w taki sposób, że maksymalna wartość tego wskaźnika w ramach rozpatrywanego modelu może wynosić 100 (minimalna 0).

Tabela 6.15 pokazuje miary jakości (zmienne niezależne), które w skali od 0 do 100 uzyskały ważność na poziomie co najmniej 50 w co najmniej jednym z 4 modeli jakości:

- **EN bin** - model jakości z dychotomiczną (binarną) zmienną zależną w angielskiej Wikipedii.
- **EN nom** - model jakości z nominalną zmienną zależną w angielskiej Wikipedii.
- **RU bin** - model jakości z dychotomiczną (binarną) zmienną zależną w rosyjskiej Wikipedii.
- **RU nom** - model jakości z nominalną zmienną zależną w rosyjskiej Wikipedii.

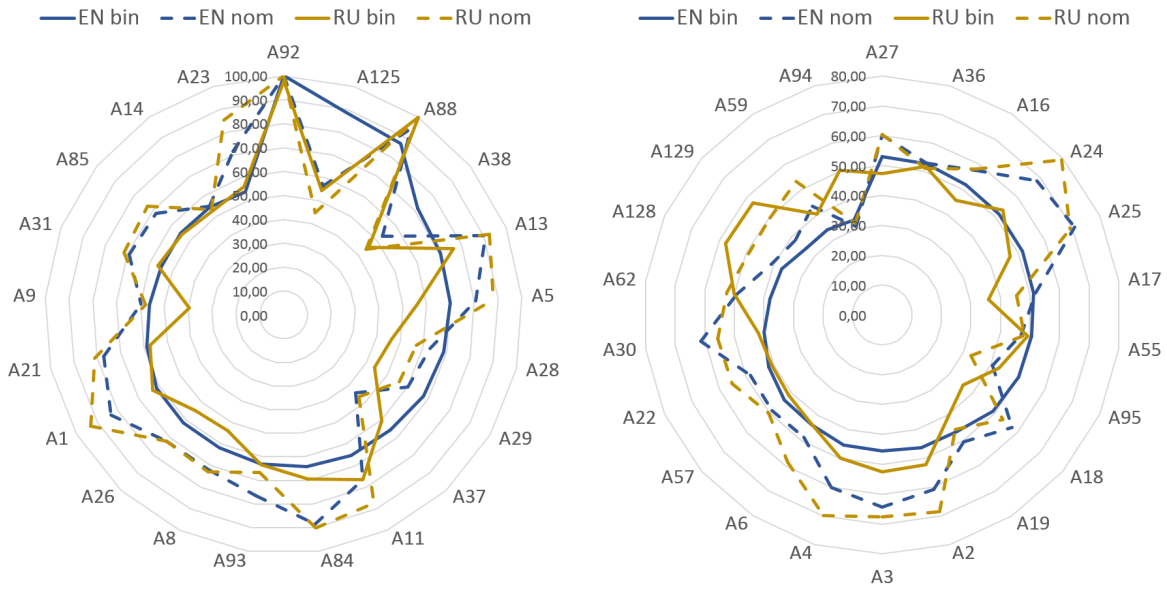
Tabela 6.15. Najważniejsze miary w modelach predykcji jakości w angielskiej (EN) lub rosyjskiej (RU) Wikipedii przy użyciu dychotomicznej (bin) lub nominalnej (nom) zmiennej zależnej z wykorzystaniem algorytmu RandomForest. Źródło: obliczenia własne przy użyciu WEKA.

Miary jakości		Model jakości			
Skrót	**Opis**	**EN bin**	**EN nom**	**RU bin**	**RU nom**
A92	Gęstość referencji	100,00	100,00	97,83	100,00
A125	Długość abstraktu na stronie dyskusji	88,87	56,44	54,56	44,80
A88	Liczba wszystkich referencji w tekście	86,73	94,55	100,00	99,97
A38	Linki przychodzące z przestrzeni nazw ns2	71,83	52,79	45,44	44,20
A13	Liczba unikatowych szablonów	70,49	90,67	76,25	92,63
A5	Liczba obrazków	69,81	80,54	55,97	87,91
A28	Długość abstraktu bez szumu w bajtach	68,78	62,89	46,31	57,08
A29	Długość abstraktu bez szumu według liczby znaków	67,71	60,18	43,93	56,02
A37	Linki przychodzące z przestrzeni nazw ns1	65,54	44,47	60,30	46,53
A11	Liczba wszystkich szablonów w tekście	65,31	76,58	76,57	87,33
A84	Suma odwiedzin za ostatni rok	64,05	88,47	69,41	90,21
A93	Długość kodu referencji	63,02	76,51	63,45	66,54
A8	Liczba sekcji	61,76	72,07	53,69	72,75
A26	Długość abstraktu w bajtach	61,57	72,12	54,34	72,20
A1	Długość artykułu w bajtach	61,19	83,54	63,12	93,09
A21	Liczba linków wewnętrznych w tekście	58,75	77,28	57,27	81,02
A9	Liczba sekcji z wyjątkiem źródłowych	56,20	59,89	39,59	58,05
A31	Maksymalna długość sekcji	55,01	69,26	56,51	71,75
A85	Mediana odwiedzin w ciągu ostatnich 90 dni	54,94	68,24	54,34	72,96
A14	Liczba szablonów 1-go poziomu w tekście	54,63	55,05	52,82	53,47
A23	Długość tekstu bez referencji w bajtach	53,79	71,75	56,18	85,27
A27	Długość abstraktu według liczby znaków	53,03	60,09	47,29	60,47
A36	Linki przychodzące z przestrzeni nazw ns0	52,61	52,91	51,84	51,08
A16	Długość szablonów 1-go poziomu w tekście	51,77	57,60	45,66	58,02
A24	Długość tekstu bez referencji według liczby znaków	51,66	68,50	53,47	79,42
A25	Czas ostatniej zmiany artykułu	51,58	70,93	47,07	69,20
A17	Długość szablonów 1-go poziomu w abstrakcie	51,28	51,77	35,90	45,41
A55	Linki przychodzące z artykułów Wikipedii	50,44	46,92	49,13	48,17
A95	Średnia liczba wypełnonych parametrów w szablonach referencji	50,13	40,50	42,95	32,71
A18	Maksymalna długość szablonu w tekście	49,26	57,61	35,68	53,53
A19	Maksymalna długość szablonu w abstrakcie	46,36	50,59	40,67	45,41
A2	Długość artykułu według liczby znaków	46,28	60,85	52,17	68,57
A3	Długość artykułu bez szumu w bajtach	45,52	64,20	52,49	67,47

A4	Długość artykułu według liczby znaków	45,48	60,19	49,89	70,02
A6	Liczba obrazków z kodu wiki w tekście	43,73	48,68	43,49	58,56
A57	Czas ostatniego odświeżenia linków	43,46	48,85	41,21	50,32
A22	Liczba linków wewnętrznych w abstrakcie	41,75	48,44	40,67	55,23
A30	Minimalna długość sekcji	39,95	61,28	41,97	55,71
A62	Data utworzenia artykułu	37,86	49,20	50,11	52,50
A128	Czas ostatniego odświeżenia na stronie dyskusji	36,98	40,94	57,70	49,20
A129	Czas ostatniego odświeżenia linków na stronie dyskusji	33,93	38,36	57,38	49,65
A59	Długość kodu źródłowego	33,93	43,31	40,24	53,20
A94	Liczba referencji ze specjalnym szablonem	33,09	30,94	50,54	32,04

Wyniki analizy ważności miar pokazują, że dla wszystkich modeli bardzo istotnym są miary związane z wiarygodnością, a w szczególności miary dotycząca gęstości referencji oraz absolutna liczba wszystkich referencji w artykule. Pierwsza z nich jest najważniejsza w 3 z 4 rozpatrywanych. Inna ważna miara dotyczy popytu na informację - suma odwiedzin za ostatni rok. Szczególnie dotyczy to modeli z nominalnej zmienną zależną dla obu wersji językowych. Niektóre miary dotyczące kompletności również wykazały wysoką ważność w modelach jakości: długość artykułu w bajtach, liczba unikatowych szablonów.

Wykres radarowy na rysunku 6.2 pokazuje różnice pomiędzy czterema rozpatrywanymi modelami jakości.



Rysunek 6.2. Ważność wybranych miar w modelach predykcji jakości w angielskiej (EN) lub rosyjskiej (RU) Wikipedii przy użyciu dychotomicznej (bin) lub nominalnej (nom) zmiennej zależnej z wykorzystaniem algorytmu RandomForest.

Źródło: Opracowanie własne.

6.6 Wykorzystanie modeli do predykcji jakości artykułów

Oceny jakości, nadawane artykułom przez użytkowników Wikipedii, mogą być przydatne dla porównywania wersji językowych pomiędzy sobą. Jednak tu mogą się pojawić problemy, wynikające głównie z dużej liczby nieocenionych artykułów oraz różnic w standardach ocen między wersjami językowymi (m.in. liczba klas jakości).

Wykorzystując zbudowane modele jakości można rozwiązać pierwszy z przedstawionych problemów. Jednak zbudować takich modeli oddzielnie dla każdej wersji językowej może być wyzwaniem: niektóre wersje językowe Wikipedii posiadają małą liczbę ocenionych artykułów, co może mieć wpływ na możliwość dopasowania modeli.

Nawet jeżeli się uda zbudować modele dla każdej wersji językowej oddzielnie, powstaje drugi ważny problem - różnice w standardach ocen pomiędzy wersjami językowymi Wikipedii. W rozdziale nr 4 „Metody określenia jakości artykułów Wikipedii" pokazano, że każda wersja językowa może mieć swobodę co do definiowania liczebności, kryteriów oraz nazw klas jakości.

W celu unifikacji klas jakości w każdej z rozpatrywanych wersji językowych oraz możliwości porównywania artykułów między sobą, następnie będą używane modele predykcji jakości zbudowane na podstawie najbardziej rozwiniętych z rozpatrywanych wersji językowych - angielskiej oraz rosyjskiej. Każda z tych wersji posiada odrębny system oceny jakości (6-7 klas jakości) oraz stosunkowo dużą liczbę ocenionych artykułów (w porównaniu do innych rozpatrywanych wersji językowych). Ocenione w ten sam sposób różne wersję językowe artykułu na wybrany temat, mogą być później porównywane pomiędzy sobą pod kątem jakości, ponieważ artykuły będą oceniony według jednego zunifikowanego standardu (angielskiej lub rosyjskiej Wikipedii). Dodatkowo, wykorzystanie modeli z różnych wersji językowych w modelach pozwoli na otrzymanie dodatkowych miar dla każdego rozpatrywanego artykułu zgodnie z różnymi systemami klasyfikacji jakości (z różnymi standardami). Te miary pozwolą na określenie współzależności pomiędzy oceną artykułu zgodnie z ogólnie przyjętymi standardami określonej wersji językowej Wikipedii a jakością odrębnych danych, które one zawierają.

Przy wykorzystaniu modeli jakości artykułów zbudowanych dla innych wersji językowych Wikipedii należy wziąć pod uwagę różnice w niektórych miarach jakości, które wynikają z specyficznych dla danego języka cech oraz reguł, według których działa konkretna społeczność językowa. Na przykład, jeżeli chcemy wykorzystać długość artykułów w bajtach, należy brać pod uwagę, że litera łacińskiego alfabetu zajmować będzie 1 bajt, a litery cyrylicy będą mieć zazwyczaj 2 bajty. To oznacza, że nawet wizualnie podobne objętościowo teksty z angielskiej oraz

białoruskiej wersji mogą mieć znaczne różnice przy obliczeniu długości w bajtach. Inne różnice mogą wynikać z różnych standardów oceniania artykułów. Na przykład, jeżeli jedna wersja językowa Wikipedii szczególnie zwraca uwagę na określoną liczbę referencji, wówczas druga wersja może mieć inny próg do nadania artykułowi podobnej oceny jakości.

Dodatkową kwestią są miary popularności. Jeżeli wersja językowa jest mniej rozwinięta (posiada stosunkowo małą liczbę artykułów), to nawet istotne dla tej wersji językowej artykuły będą mieć mniej unikatowych autorów, mniej linków do tego artykułu z innych stron Wikipedii. Do miar popularności należy również liczba odwiedzin stron w określonym czasie. Można się spodziewać, że większe wersje językowe będą odwiedzane średnio częściej.

W związku z powyższym wartości miar przed zbudowaniem modeli na podstawie angielskiej oraz rosyjskiej Wikipedii, wszystkie miary artykułów zostaną przekształcone przy użyciu normalizacji min-max. Przy tym, największą wartością (*max*) będzie równa wartości mediany maksymalnych wartości tej miary dla *n* najlepszych artykułów w danej wersji językowej Wikipedii. Zmienna *n* wskazuje na liczbę artykułów najwyższej jakości w danym języku. Na przykład dla angielskiej Wikipedii ta liczba wynosi 5344. Wcześniejsze badania pokazały, że w większości przypadków wyższa jakość artykułu jest skorelowana z wyższymi wartościami różnych miar tego artykułu. Z drugiej strony, trzeba również uwzględnić inne sytuacje, kiedy największe wartości miar nie muszą wskazywać na lepszą jakość artykułu (korelacja ujemna). Mediana została wybrana jako wartość środkowa, ponieważ pozwala uniknąć dużego wpływu skrajnych wartości.

Najpierw zostały zbudowane modele dla angielskiej oraz rosyjskiej wersji przy użyciu znormalizowanych wartości miar. Modeli były zbudowane dla dychotomicznej oraz kategorialnej zmiennej zależnej, podobnie jak to było zrobione w poprzedniej sekcji. Do budowania modeli został wykorzystany algorytm lasu losowego. W związku z tym, że skalowanie liniowe nie wpływa na różnice pomiędzy artykułami w różnych klasach, wszystkie 4 zbudowane modele na wartościach znormalizowanych wykazały precyzję podobną do modeli budowanych na wartościach absolutnych.

Do wykorzystania tych modeli zostały dobrane artykuły na podobny temat. Jeżeli rozpatrywać artykuły o polskich miastach we wszystkich 5 wersji językowych, to do oceny może być wybrano tylko 200 artykułów. W celu zwiększenia próby, została wykluczona najmniejsza z rozpatrywanych wersji językowych Wikipedii (białoruska) i została przeprowadzona analiza dla 4 wersji językowych (angielska, rosyjska, polska, ukraińska). W tym przypadku liczba artykułów o polskich miastach wyniosła 901. Sposób doboru tych artykułów jest opisany w następnym

rozdziale. Dla każdego artykułu zostały wyekstrahowane te same miary, które były wcześniej wykorzystane do budowania modeli jakości. Następnie wartości miar, zostały znormalizowane, zgodnie z regułami opisanymi wyżej w danej sekcji.

Najpierw do oceny artykułów został użyty modeli oceny jakości angielskiej Wikipedii. Zostały użyte modele z dychotomiczną oraz modele z nominalną zmienną zależną mają. Wyniki pokazano w tabeli 6.16.

Tabela 6.16. Liczba artykułów o polskich miastach w 4 wersjach językowych ocenionych przy pomocy modeli jakości angielskiej Wikipedii z użyciem dychotomicznej oraz kategorialnej zmiennej zależnej.

Wersja językowa	Jakość kategorialna						Jakość dychotomiczna	
	FA	GA	B	C	Start	Stub	Kompletne	NieKompletne
EN - angielska	7	3	64	62	353	412	12	889
PL - polska	1	11	328	492	69	0	21	880
RU - rosyjska	0	0	8	15	121	757	0	901
UK - ukraińska	0	0	4	25	118	754	0	901

Źródło: Opracowanie własne.

Następnie te same artykuły zostały ocenione przez modele jakości rosyjskiej Wikipedii. Wyniki przedstawia tabela 6.17.

Tabela 6.17. Liczba artykułów o polskich miastach w 4 wersjach językowych ocenionych przy pomocy modeli jakości rosyjskiej Wikipedii z użyciem dychotomicznej oraz kategorialnej zmiennej zależnej.

Wersja językowa	Jakość kategorialna							Jakość dychotomiczna	
	FA	GA	SA	IV	III	II	I	Kompletne	NieKompletne
EN - angielska	12	22	19	426	231	148	43	11	890
PL - polska	59	37	3	0	5	366	431	24	877
RU - rosyjska	0	2	1	768	54	67	9	0	901
UK - ukraińska	0	1	2	743	81	62	12	0	901

Źródło: Opracowanie własne.

Wyniki predykcji jakości artykułów są niejednoznaczne. Z jednej strony, przy użyciu dychotomicznej zmiennej zależnej modeli angielskiej oraz rosyjskiej Wikipedii pokazują największą liczbę artykułów wysokiej jakości w polskojęzycznej wersji (w kategorii „Kompletne”). Z drugiej strony, przy użyciu nominalnej zmiennej zależnej modelu angielskiej Wikipedii, większą liczbę

najlepszych artykułów (z klasy FA) posiada angielska wersja. Jeżeli dodatkowo uwzględnimy artykułu z innej wysokiej klasy jakości (GA), to sumaryczna liczba artykułów będzie większa w polskojęzycznej wersji. Inny jest wynik przy użyciu modelu rosyjskiej Wikipedii z nominalną zmienną zależną, który określił znacznie więcej artykułów jako najlepsze (klasa FA) dla polskiej Wikipedii. Dodatkowo, model rosyjskiej Wikipedii wykrył więcej artykułów najwyższej jakości dla angielskiej wersji.

Otrzymane oceny używane w następnych rozdziałach jako dodatkowe miary, w celu określenia, na ile ważne są oceny nadawane przez użytkowników Wikipedii dla artykułów w stosunku do jakości odrębnych informacji zawartych w tych artykułach.

W tym celu klasy jakości zostaną przekształcone w liczby w taki sposób, żeby lepsza jakość miała większą wartość. Tabela 6.18 pokazuje odpowiednie wartości liczbowe dla poszczególnych klas jakości w angielskiej oraz rosyjskiej Wikipedii. Należy zaznaczyć, że w związku z tym, że klasa SA w rosyjskiej Wikipedii nie ma odpowiednika w angielskiej wersji językowej, a jednocześnie artykuły do tej klasy przypisywany są według procedury podobnej do najwyższych klas jakości (FA, GA), to wartość liczbowa tej klasy musi być niższa od tych najlepszych oraz jednocześnie wyższa od pozostałych klas, gdzie ocena może być wystawiana indywidualnie przez dowolnego użytkownika.

Tabela 6.18. Wartości liczbowe przypisane poszczególnym klasom jakości w angielskiej (EN) oraz rosyjskiej (RU) Wikipedii.

Klasa EN	Wartość
FA	7
GA	6
B	4
C	3
Start	2
Stub	1

Klasa RU	Wartosć
FA	7
GA	6
SA	5
IV	4
III	3
II	2
I	1

Źródło: Opracowanie własne.

W przypadku modeli jakości z dychotomiczną zmienną zależną, „Kompletne” będą oznaczane wartością 1, a „NieKompletne” - wartością 0. W następnych rozdziałach przy budowaniu modelu jakości infoboksu wyniki ocen modeli z dychotomiczną oraz nominalną zmienną zależną były dodatkowo normalizowane.

W związku z powyższym, dla każdej z rozpatrywanych wersji artykułu będą dodane 4 dodatkowe miary:

- J_1 - jakość artykułu według modelu angielskiej Wikipedii z kategorialną zmienną zależną w skali od 1 do 7,
- J_2 - jakość artykułu według modelu angielskiej Wikipedii z dychotomiczną zmienną zależną w skali od 0 do 1,
- J_3 - jakość artykułu według modelu rosyjskiej Wikipedii z kategorialną zmienną zależną w skali od 1 do 7,
- J_4 - jakość artykułu według modelu rosyjskiej Wikipedii z dychotomiczną zmienną zależną w skali od 0 do 1.

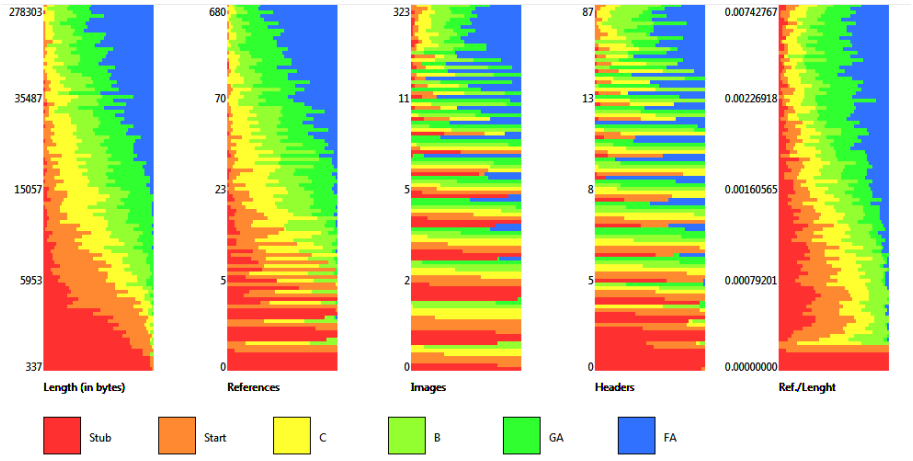
6.7 Miara syntetyczna

Duża część tej sekcji jest napisana na podstawie wcześniejszych badań (Lewoniewski i Węcel, 2017; Lewoniewski, Węcel i Abramowicz, 2017b).

Innym sposobem mierzenia jakości jest stosowanie miary syntetycznej (czy też wskaźnika syntetycznego). To jest wartość, która jest wynikiem połączenia innych oszacowań poszczególnych miar (kryteriów czy cech). Na przykład taki wskaźnik może reprezentować sumę punktów ocen ekspertów. Miara syntetyczna jest często stosowana w gospodarce i może opisać m.in. obiekt gospodarczy czy system gospodarczy jako całość (Aganbegjan, 2012; Podshivalenko, 2010; Shvecov, 2011).

Miara syntetyczna może pozwolić na ocenę artykułów Wikipedii w postaci zmiennej ciągłej (Lewoniewski, Węcel i Abramowicz, 2017b). Jak było wspomniane w rozdziale nr 4 „Metody określenia jakości artykułów Wikipedii", każda wersja językowa może mieć swój system ocen artykułów i swoją liczbę klas jakości. W dalszej części rozważań skupiono się na największej wersji językowej Wikipedii - angielskiej.

Przed zbudowaniem miary syntetycznej przeprowadzono wstępną analizę wybranych miar jakości. Z 6 klas jakości zostało losowo wybrano 1000 artykułów (z każdej klasy) i dla każdego z nich zostało wyekstrahowane pięć miar: długość artykułu, liczba referencji, liczba obrazków, liczba sekcji, gęstość referencji. Wyniki prac (Lewoniewski i in., 2016; Warncke-wang i in., 2013; Węcel i Lewoniewski, 2015) oraz własne eksperymenty pokazały, że te miary znajdują się wśród najważniejszych wykorzystywanych w modelach oceny jakości. Rysunek 6.3 pokazuje rozkład



Rysunek 6.3. Rozkład wybranych miar w artykułach każdej klasy jakości w angielskiej Wikipedii (FA - najwyższa klasa, Stub - najniższa).

Źródło: Obliczenia własne.

wartości miar wraz udziałem artykułów z poszczególnych klas jakości. Można zauważyć, że im większa wartość miary (odłożona na osi pionowej), tym większy jest udział artykułów o wysokiej jakości. Przykładowo, jeżeli bierzemy pod uwagę miarę długość artykułów, można się spodziewać, że im dłuższy jest artykuł, tym większe prawdopodobieństwo posiadania przez niego wyższej klasy jakości.

Pokazane zależności prowadzą do ogólnego wniosku, że większą wartość skumulowanej miary syntetycznej, która łączy najważniejsze miary, zostanie przypisana dla bardziej rozwiniętych artykułów (czyli tych artykułów, które posiadają wyższą ocenę jakości).

Dodatkowy aspekt, który był wzięty pod uwagę - różnica w kryteriach oceny artykułów w każdej wersji językowej Wikipedii. Na przykład w konkretnej wersji językowej użytkownicy mogą przydzielać większą uwagę do liczby referencji niż do liczby obrazków przy decydowaniu o nadaniu wysokie oceny za jakość. Poza tym wystarczająca liczba referencji do nadania określonej oceny jakości też jest określana różnie w zależności od wersji językowej. Dla tego oddzielnie dla każdej wersji językowej należy zbadać wybrane miary jakości najlepszych (wzorcowych) artykułów.

Najwyższa klasa jakości (FA - w angielskiej, ANM - w polskiej) jest obecna w każdej z rozpatrywanych wersji językowych. Artykuły mogą otrzymać taką ocenę z czasem, kiedy zawartość będzie odpowiadała określonym kryteriom: np. będzie zawierała wystarczająca liczbę referencji, obrazków, sekcji etc. Można powiedzieć, że artykuł dąży do określonego progu, w którym może dostać najwyższą ocenę. Taki próg może mieć każda z rozpatrywanych miar. W celu ob-

liczenia tych progów, zostały wyekstrahowane miary wszystkich artykułów z najwyższej klasy z każdej z rozpatrywanych wersji językowych. Następnie została obliczona mediana dla każdej miary w każdym języku. Wyniki pokazane są w tabeli 6.19.

Tabela 6.19. Mediany wartości miar w najwyższej klasie jakości w różnych językach Wikipedii.

Język	Długość	Referencje	Obrazki	Sekcje	Ref./Długość
BE	198 365	210	36	27	0,001106
EN	49 038	115	13	14	0,002364
PL	59 672	96	17	17	0,001663
RU	139 415	163	24	22	0,001169
UK	82 371,5	40,5	24,5	21	0,000491

Źródło: Obliczenia własne.

Obliczone mediany miar będą stanowiły podstawę do normalizacji tych miar przed obliczeniem wskaźnika syntetycznego. Innymi słowy, ta mediana wartości stanowi próg, który pokazuje stopień rozwoju artykułu według określonej miary. Przy tym, jeżeli wartość wybranej miary jest wyższa niż odpowiednia mediana (czy odpowiedni próg), to wartość znormalizowana będzie równa 1. Innymi słowy, znormalizowana miara *i* obliczana jest według wzoru:

$$\begin{cases} \frac{m_i}{p_i}, m_i < p_i \\ 1, m_i \geq p_i \end{cases} \tag{6.1}$$

gdzie m_i to absolutna wartość miary *i* oraz p_i to mediana miary *i* w najwyższej klasie jakości danej wersji językowej

Zakładamy, że wszystkie miary mogą mieć podobny wpływ na jakość, a w związku z tym muszą mieć równy wpływ na wartość wskaźnika syntetycznego. Najpierw należy obliczyć średnią znormalizowanych miar (SZM) według następującego wzoru:

$$SZM = \frac{1}{c} \sum_{i=1}^{c} \tag{6.2}$$

gdzie nm_i to znormalizowana miara *i* oraz *c* pokazuje liczbę miar.

Następnie bierzemy pod uwagę liczbę szablonów wad jakości (*SWJ*) w rozpatrywanym artykule (jeśli istniały).

Biorąc pod uwagę powyższe założenia, wskaźnik syntetyczny, określający jakość artykułów, będzie liczony na podstawie wzoru:

Tabela 6.20. Zaokrąglone wartości wskaźnika syntetycznego dla artykułów o polskich miastach w 4 wersjach językowych Wikipedii.

Wersja językowa	Zaokrąglona wartość wskaźnika syntetycznego										
	0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
EN - angielska	0	142	249	233	112	100	25	19	12	5	4
PL - polska	0	0	0	2	59	272	307	132	69	35	25
RU - rosyjska	0	594	207	67	15	13	3	1	1	0	0
UK - ukraińska	17	414	318	81	42	12	9	5	1	2	0

Źródło: Obliczenia własne.

$$Jakosc = SZM - SZM \cdot 0,05 \cdot SWJ \tag{6.3}$$

gdzie *SZM* średnią znormalizowanych miar oraz *SWJ* pokazuje liczbę szablonów wad jakości.

W tym przypadku wartość wagi 0,05 dla liczby szablonów o lukach jakości dobrana została na podstawie badań (Anderka, 2013) oraz własnych obserwacji, które pokazały, że artykuły które posiadają co najmniej jeden taki szablon nie mogą mieć wartości miary syntetycznej wyższej niż mediana liczby punktów najlepszych artykułów Wikipedii w danej wersji językowej obliczonej według wzoru 6.2. Na przykład, jeżeli artykuł posiada wszystkie wartości miar *i* wyższe niż odpowiednie mediany wartości miar p_i, to wartość *SZM* będzie maksymalnej (czyli 1). W przypadku jeżeli artykuł posiada dwa szablony wskazujący na luki w jakości, to zgodnie ze wzorem 6.3 ta wartość będzie obniżona o 0.1, co w wyniku da wartość 0.9 za jakość.

Zmienna ciągła pozwala na utworzenie dowolnej liczby klas (ocen) jakości. Na przykład, jeżeli wartości wskaźnika syntetycznego zaokrąglić do części dziesiątych, możemy otrzymać 11 ocen z zakresu: 0, 0.1, 0.2, ... , 0.9, 1. Stosując taką skalę ocen, zostały ocenione wcześnie rozpatrywane artykuły o polskich miastach w 4 wersjach językowych Wikipedii. Wyniki oceny pokazane w tabele 6.20

Wyniki analizy pokazują, że nawet przy zwiększonej liczbie ocen w porównaniu do standardowej liczby ocen w różnych wersja językowych Wikipedii (por. tabele 6.16 oraz 6.17), otrzymanych po zaokrągleniu wartości ciągłej, wskaźnik syntetyczny pokazuje znacznie większą przewagę artykułów lepszej jakości w polskojęzycznej wersji językowej.

Dodatkowe eksperymenty pokazały, iż wskaźnik syntetyczny jako dodatkowy predyktor może zwiększyć precyzję w modelach klasyfikacyjnych. Na przykład, jeżeli w angielskiej Wikipe-

dii do modelu klasyfikacyjnego z kategorialną zmienną zależną wprowadzić dodatkowo wskaźnik syntetyczny jako zmienną niezależną, to precyzja modelu się zwiększa, a sam wskaźnik ma najwyższą ważność jako predyktor.

Znormalizowane miary oraz sam wskaźnik syntetyczny używane w następnych rozdziałach jako dodatkowe miary w budowaniu modeli jakości infoboksów, które będą opisane w następnej sekcji:

- J_5 - ocena jakości artykułu z wykorzystaniem miary syntetycznej,
- A_{134} - długość artykułu znormalizowana do miary syntetycznej,
- A_{135} - liczba referencji artykułu znormalizowana do miary syntetycznej,
- A_{136} - liczba obrazków artykułu znormalizowana do miary syntetycznej,
- A_{137} - liczba sekcji artykułu znormalizowana do miary syntetycznej,
- A_{138} - gęstość referencji artykułu znormalizowana do miary syntetycznej.

6.8 Podsumowanie

Na podstawie ponad 130 miar zostały zbudowane modele jakości przy wykorzystaniu ponad 20 algorytmów. Największą precyzję wykazał algorytm lasu losowego (Random Forest), który został wykorzystany w kolejnych rozdziałach rozprawy do budowania modeli jakości infoboksów oraz identyfikacji najważniejszych miar jakości.

Modele jakości artykułu były zbudowane z wykorzystaniem dychotomicznej oraz nominalnej zmiennej zależnej. Większą precyzję można osiągnąć przy użyciu dychotomicznej zmiennej zależnej, kiedy jakość jest modelowana jako prawdopodobieństwo przynależności do jednej z dwóch kategorii: „Kompletne” i „NieKompletne”. Przy nominalnej zmiennej zależnej liczba kategorii była większa i zależała od wersji językowej. Wśród najlepszych predyktorów w tych modelach były miary dotyczące wiarygodności, popytu oraz kompletności artykułów Wikipedii.

Modele zostały zbudowane na podstawie angielskiej oraz rosyjskiej Wikipedii. Te modele zostały dodatkowo użyte do oceny jakości wybranych artykułów na temat polskich miast. Wykorzystanie znormalizowanych miar pozwoliło na wykorzystanie tych modeli w różnych wersjach językowych.

Wyniki analizy pokazały, że istnieją różnicę pomiędzy modelami jakości w różnych wersjach językowych przy wykorzystaniu podobnych zmiennych zależnych. To zostało pokazano m.in. przy analizie istotności miar w tych modelach. W przypadku analizy różnych wersji językowych

artykułów opisujących polskie miasta, modele z dychotomiczną zmienną zależną obu wersji językowych pokazały podobne wyniki oceny jakości - największa liczba artykułów z najwyższą jakością posiadała polska Wikipedia. Przy wykorzystaniu nominalnej zmiennej zależnej można było zaobserwować większe różnice pomiędzy liczbami artykułów najwyższej jakości. W tym przypadku model angielskiej Wikipedii oznaczył więcej artykułów najwyższą klasą FA w wersji angielskiej.

Różnice w modelach jakości wynikają przede wszystkim z różnić pomiędzy standardami ocen w każdej wersji językowej. Modele, zbudowane na podstawie angielskiej oraz rosyjskiej Wikipedii będą wykorzystane jako dodatkowe miary w następnych rozdziałach w celu określenia ważności ocen, nadanych przez użytkowników Wikipedii a jakością określonych informacji w tych artykułach.

Dodatkowo w tym rozdziale został wprowadzony wskaźnik syntetyczny, który pozwala na ocenę jakości artykułów przy użyciu wartości ciągłej od 0 do 1. Ta zmienna pozwala na utworzenie dowolnej liczby klas (ocen) jakości. Ten wskaźnik pokazał większą liczbę artykułów o polskich miastach z wysoką jakością w polskojęzycznej Wikipedii. Dodatkowo, wskaźnik syntetyczny posiada wysoką ważność, jeżeli uwzględnić go jako dodatkowy predyktor we wcześniej zbudowanych modelach klasyfikacyjnych. W związku z tym, wskaźnik syntetyczny będzie również używany w następnych rozdziałach jako dodatkowy predyktor w modelach oceny określonych informacji w artykułach Wikipedii.

Rozdział 7

Miary oraz wymiary jakości infoboksów

W niniejszym rozdziale przedstawione dodatkowe miary jakości, które dotyczą poszczególnej części artykułów Wikipedii - infoboksów. Zazwyczaj te infoboksy które prezentują najważniejsze informacje na określony temat. Również w tym rozdziale przedstawione sposoby ekstrakcji oraz obliczania miar infoboksów.

Dla niektórych tematów artykułów przedstawiona analiza miar w ramach infoboksów. W celu odwołania się do miar jakości infoboksów wprowadzono oznaczenie I_x (lub *Ix*), gdzie *x* - to indeks (numer) miary. Wszystkie miary infoboksów zostały przypisane do określonych wymiarów jakości: kompletność, wiarygodność, aktualność oraz relewancja.

Obszerne materiału tego rozdziału opracowane na podstawie wcześniejszych badań (Lewoniewski, 2017a).

7.1 Wprowadzenie

Infoboksy używane w podobnych artykułach Wikipedii w celu zapewnienia spójności prezentacji przy użyciu wspólnego formatu (Yu, 2011). Infoboksy w niektórych aspektach są one porównywalne z tabelami danych. Jakość danych może zależeć od różnych wymiarów (czy charakterystyk), takich jak kompletność, aktualność, dokładność oraz inne (Heinrich i Klier, 2015). Dodatkowo, mogą one zawierać obrazki, referencję, szablony, odnośniki do innych stron oraz inne elementy. Innymi słowy, infoboks niesie za sobą informacje a nie tylko dane - ponieważ stoją za tym struktury semantyczne. Dane w tych infoboksach mogą być dodatkowo zmieniane przez użytkowników, np. w celu aktualizacji danych. W związku z tym, jakość infoboksów może

być charakteryzowana przez różne wymiary jakości: aktualność, wiarygodność, kompletność oraz inne.

Wstępne analizy pokazały, że artykuły posiadające wyróżnienie przez użytkowników za wysoką jakość nie zawsze posiadają informacje najwyższej jakości w umieszczonym w artykule infoboksie.

Mechanizm wyszukiwania w Wikipedii pozwala znaleźć wszystkie artykuły, które zawierają określony infoboks. W celu znalezienia artykułów na określony temat można użyć odpowiednią nazwę (czy nazwy) infoboksów w różnych językach. Np. w angielskiej Wikipedii artykuły o firmach zazwyczaj używają infoboks o nazwie „Infobox company", w polskiej - „Przedsiębiorstwo infobox". W związku z tym, na podstawie infoboksów można otrzymać listę nazw artykułów na określony temat w każdej wersji językowej. Z drugiej strony, niektóre artykuły w określonej wersji językowej mogą nie posiadać infoboksu. W tym przypadku można użyć powiązania semantyczne z innymi wersjami językowym, które te infoboksy posiadają.

Warto zaznaczyć, iż wybór infoboksów jako kryterium wyszukiwawczego artykułów w ramach tej rozprawy był celowy. W Wikipedii artykuły są również klasyfikowane według systemu kategorii, jednak przypisanie kategorii wymaga od edytora umieszczenia w treści artykuły dodatkowego kodu. Artykuły mogą być przypisywane do szerokiego zakresu kategorii, nie zawsze tematycznie powiązanych z ich treścią. Dodatkowo kategorie są budowane w taki sposób, że przechodząc do podkategorii można przejść do artykułów innego rodzaju: np. w kategorii „Gry komputerowe" można razem z podkategoriami „Gry komputerowe według platformy", „Gry komputerowe według roku wydania" znaleźć m.in. „Wydawcy gier komputerowych", „Producenci gier komputerowych" czy też „Prawo i cenzura gier komputerowych". Kategorie nie tworzą zatem taksonomii, a więc nie można wnioskować o typach artykułów.

7.2 Ekstrakcja parametrów infoboksów

Do przeprowadzenia dalszych analiz zostały wybrane artykuły z 5 rozpatrywanych wersji językowych Wikipedii z infoboksami opisujących następujące tematy: albumy, gry wideo, filmy, firmy, uniwersytety.

Tabela 7.1 pokazuje liczbę artykułów z infoboksami na określone tematy w poszczególnych wersjach językowych Wikipedii.

Tabela 7.1. Liczba artykułów z infoboksami na określone tematy w poszczególnych wersjach językowych Wikipedii.

Temat	BE	EN	PL	RU	UK
Albumy	130	137 972	22 026	14 144	6 522
Firmy	371	56 678	4 660	9 449	3 628
Filmy	212	114 727	18 654	25 615	12 879
Uniwersytety	244	20 421	2 175	2 320	1 082
Gry wideo	51	20 685	2 924	5 492	1 341

Źródło: Obliczenia własne.

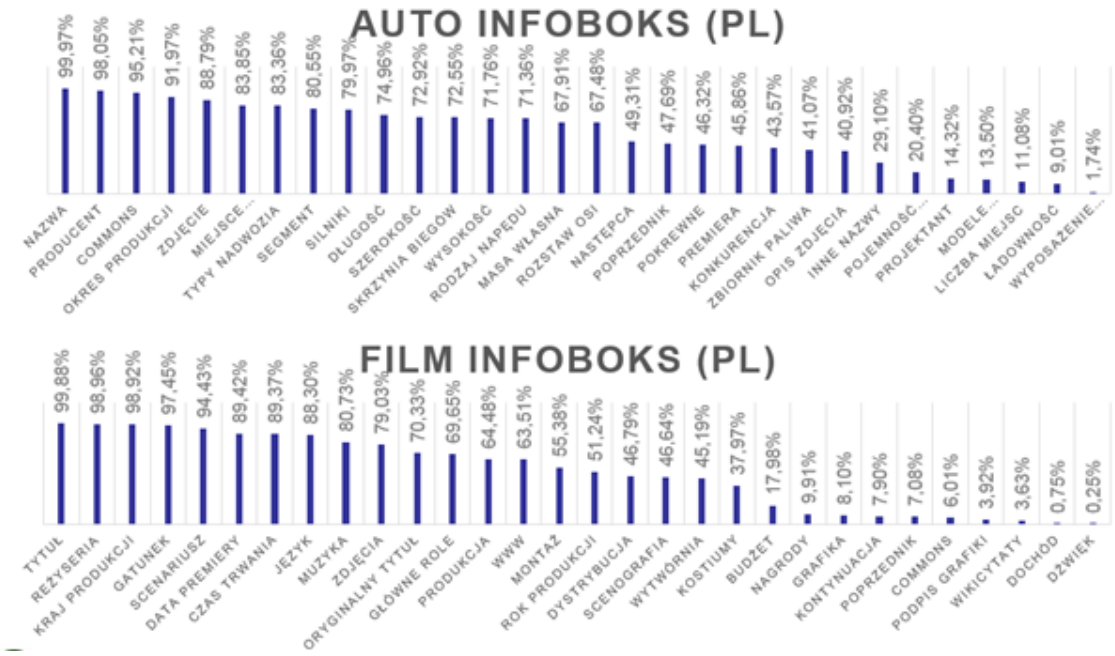
W celu ekstrakcji parametrów z infoboksów z różnych wersji językowych został przygotowany specjalny parser w języku Python. Taki program pozwala m.in. określić, w jakim stopniu autorzy Wikipedii wypełniają poszczególne parametry w infoboksach. Jeśli chodzi o polską Wikipedię i grupy produktów samochody oraz filmy, to prawie w 100% przypadków w takich infoboksach jest uzupełniany parametr „nazwa" („tytuł"). Bardziej specyficzne parametry uzupełniane są znacznie rzadziej, np. „Ładowność", "Projektant", "Dochód" może posiadać wartość w mniej niż 20% wszystkich wystąpień danego infoboksu. Szczegółowe wyniki częstości wypełniania poszczególnych parametrów dla 2 grup produktów w Polskiej Wikipedii przedstawione na rys. 7.1.

Następnie została przeprowadzona analiza list parametrów infoboksów w różnych językach. Najpierw porównujemy liczebności różnych parametrów w infoboksach tego samego typu. Wyniki przedstawione zostały w tabeli 3 i pokazują różną „kulturę" w definiowaniu dostępnych parametrów w infoboksach danego typu w czterech różnych wersjach językowych Wikipedii.

7.3 Miary jakości infoboksów

Jak już było wspomniane wcześniej, infoboks można rozpatrywać jako tabelę z danymi, która może zawierać dane różnego rodzaju, w tym referencje. Na rys. 7.2 pokazane zostały przykładowe miary jakości infoboksów, które liczone na podstawie wypełnionych parametrów w takiej tabeli oraz liczby wszystkich i unikatowych referencji.

W dalszych podsekcjach zostały omówione miary z poszczególnych wymiarów jakości infoboksów. Wymiary jakości infoboksów zostały dobrane w taki sposób, aby istniała możliwość porównywania z odpowiednimi wymiarami jakości artykułów.



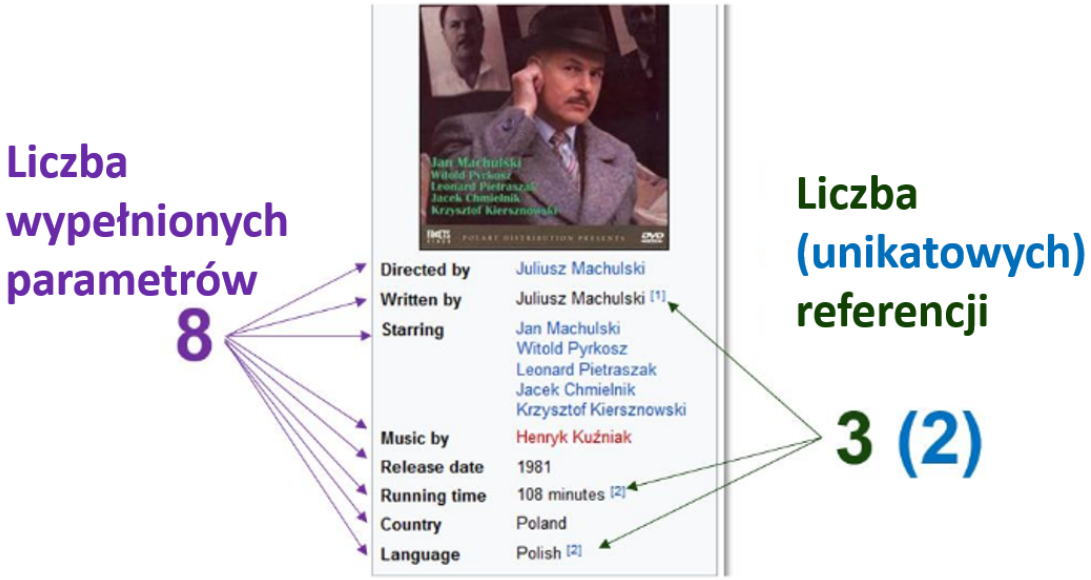
Rysunek 7.1. Częstość wypełniania parametrów infoboksów w polskiej Wikipedii.

Źródło: Obliczenia własne.

Tabela 7.2. Liczba używanych parametrów w infoboksach. Brane pod uwagę parametry, które posiadały wartości w co najmniej 5 infoboksach danego typu

Temat	DE	EN	PL	RU
Albumy	67	105	67	68
Gry wideo	40	39	33	143
Telefony kom.	50	88	26	101
Samochody	24	71	32	51
Filmy	65	253	34	84
Oprogramowanie	34	107	16	69

Źródło: Obliczenia własne.



Rysunek 7.2. Wybrane miary jakości infoboksu o filmie.

Źródło: Opracowanie własne.

7.3.1 Kompletność

Każdy parametr infoboksu może być oddzielnie wypełniany przez użytkowników Wikipedii, dlatego każdy parametr może mieć różną częstość wypełniania w ramach infoboksu określonego rodzaju. Tabela 7.3 przedstawia częstotliwość wypełnienia poszczególnych parametrów w infoboksie o firmach w różnych wersjach językowych Wikipedii. W związku z tym, ze zazwy parametrów infoboksów w różnych wersjach językowych mogą mieć różne nazwy, dla danej tabeli zostały użyte zunifikowane nazwy, które wykorzystywane są w semantycznej bazie danych DBpedia (więcej informacji o unifikacji można znaleźć w rozdziale 9 „Porównanie informacji wielojęzycznych")

Tabela 7.3. Częstotliwość wypełnienia poszczególnych parametrów w infoboksie o firmach w różnych wersjach językowych Wikipedii.

Parametr	BE	EN	PL	RU	UK
Name	.758	.978	.999	.932	.780
Industry	.617	.858	.821	.790	.675
Foundation	.705	.801	.893	.873	.520
Type	.640	.764	.660	.710	.540
Homepage	.701	.745	.813	.757	.623

Źródło: Obliczenia własne.

Kompletność infoboksu można mierzyć jako stosunek liczby wypełnionych wartościami parametrów do liczby wszystkich zdefiniowanych parametrów w infoboksie danego typu (Zaveri i in., 2016). Pierwszy sposób obliczania kompletności infoboksu polega na liczeniu liczby wypełnionych parametrów w ramach infoboksu. Ta miara będzie mieć oznaczenie I_1. Drugi sposób obliczania kompletności polega na uwzględnieniu liczby innych zdefiniowanych parametrów w ramach danego infoboksu oraz uwzględnieniu wagi każdego z wypełnionych parametrów.

$$I_2 = \frac{\sum_{i=1}^{UP} w_i}{ZP} \tag{7.1}$$

gdzie *UP* uzupełnione parametry, w_i - waga parametru *i*, *ZP* - wszystkie zdefiniowane parametry w danym infoboksie.

Wartość wagi *w* jest oparta na częstotliwości wypełniania konkretnego parametru infoboksu. Na przykład dla infobosku opisującego uniwersytety w angielskiej Wikipedii, waga parametru „city" jest równa 0,9347, ponieważ w około 93% infoboksów danego typu w danej wersji językowej ten parametr jest wypełniony.

W tabeli 7.4 przedstawione średnie wartości kompletności I_2 infoboksów na różne tematy w poszczególnych wersjach językowych Wikipedii.

Tabela 7.4. Średnie wartości kompletności I_2 infoboksów na różne tematy w poszczególnych wersjach językowych Wikipedii.

Temat	BE	EN	PL	RU	UK
Albumy	.217	.452	.341	.474	.329
Firmy	.071	.107	.153	.131	.04
Filmy	.527	.36	.402	.518	.199
Uniwersytety	.158	.099	.204	.19	.114
Gry wideo	.044	.287	.266	.043	.048

Źródło: Obliczenia własne.

Niżej opisane są inne sposoby mierzenia kompletności infoboksów, które będą używane w rozprawie:

- I_3 - liczba wszystkich wypełnionych parametrów I_1, z wyjątkiem tych, które mają wagę mniejszą niż 0.1,
- I_4 - mediana liczby wszystkich wypełnionych parametrów I_1 dla danego typu infoboksu w wybranej wersji językowej Wikipedii,

- I_5 - długość w bajtach wszystkich wartości parametrów infoboksu,
- I_6 - średnia długość wartości infoboksu: $I_6 = I_5/I_1$,
- I_7 - liczba linków do artykułów Wikipedii w wartościach infoboksu,
- I_8 - długość wszystkich wartości parametrów bez uwzględnienia kodu, opisującego referencje,
- I_9 - liczba szablonów w wartościach parametrów,
- I_{10} - liczba wszystkich wypełnionych parametrów (I_1), z wyjątkiem tych, które w 50% przypadkach mają podobne wartości,
- I_{11} - liczba wszystkich wypełnionych parametrów z uwzględnieniem wag (I_2), z wyjątkiem tych parametrów, które w co najmniej 50% przypadkach mają podobne wartości,
- I_{12} - długość w bajtach wszystkich wartości parametrów infoboksu, z wyjątkiem tych parametrów, które w co najmniej 50% przypadkach mają podobne wartości,
- I_{13} - długość w bajtach wszystkich wartości parametrów bez uwzględnienia kodu, opisującego referencje oraz z wyjątkiem tych parametrów, które w co najmniej 50% przypadkach mają podobne wartości,
- I_{14} - średnia długość wartości parametrów, które nie mają w co najmniej 50% przypadkach podobne wartości: $I_{14} = I_{12}/I_{10}$,
- I_{15} - długość wszystkich wartości parametrów infoboksu według liczby znaków,
- I_{16} - średnia długość wartości infoboksu według liczby znaków: $I_{16} = I_{15}/I_1$,
- I_{17} - długość wszystkich wartości parametrów według liczby znaków bez uwzględnienia kodu, opisującego referencje,
- I_{18} - długość według liczby znaków wszystkich wartości parametrów infoboksu, z wyjątkiem tych parametrów, które w co najmniej 50% przypadkach mają podobne wartości,
- I_{19} - długość według liczby znaków wszystkich wartości parametrów bez uwzględnienia kodu, opisującego referencje oraz z wyjątkiem tych parametrów, które w co najmniej 50% przypadkach mają podobne wartości,
- I_{20} - średnia długość według liczby znaków wartości parametrów, które nie mają w co najmniej 50% przypadkach podobne wartości: $I_{20} = I_{18}/I_{10}$,

Warto oddzielnie zwrócić uwagę na miary I10-I14 oraz I18-I20, gdzie są wykluczane parametry, które mają podobne wartości w co najmniej 50% przypadków w ramach rozpatrywanego infoboksu. W niektórych wersjach językowych, są stosowane ogólne infoboksy na określony temat, gdy inna wersja językowa ma szczegółowy podział w zależności od cech podmiotu

Tabela 7.5. Średnia liczba referencji I_{21} w infoboksach na różne tematy w poszczególnych wersjach językowych Wikipedii.

Temat	BE	EN	PL	RU	UK
Albumy	.22	.153	1.002	.641	.187
Firmy	.386	.649	.352	.56	.459
Filmy	.553	.441	.403	.177	.316
Uniwersytety	.236	.762	.363	.329	.337
Gry wideo	1.8	.807	1.944	.874	.641

Źródło: Obliczenia własne.

(np. przynależność do państwa - miasta). Uniwersalność takich szablonów wymaga zwiększenia liczby parametrów, w celu doprecyzowania cech opisywanego podmiotu. Na przykład, jeżeli istnieje infoboks opisujący miasto Polski, to nie ma tam oddzielnego parametru, w którym można wpisać państwo do którego należy to miasto. Innymi słowy, szczególne przypadki infoboksów domyślnie zawierają uzupełnione parametry, wówczas gdy w uniwersalnych odpowiednikach podobny parametr należy uzupełnić. Dlatego dla obliczenia niektórych miar nie brane były pod uwagę parametry, które w większości (ponad 50 proc) infoboksów w rozpatrywanych artykułach zawierały identyczną wartość.

7.3.2 Wiarygodność

Jednym ze wygodnych sposobów weryfikacji wiarygodności informacji w Wikipedii jest sprawdzenie źródeł. W związku z tym do mierzenia wiarygodności infoboksów można użyć takie mary jak: liczba referencji (I_{21}), liczba unikatowych referencji (I_{22}) oraz stosunek liczby referencji do liczby wypełnionych parametrów w infoboksie (I_1) liczonej wg wzoru:

$$I_{23} = \frac{I_{21}}{I_1} \tag{7.2}$$

gdzie I_{21} – liczba referencji, I_1 – liczba uzupełnionych parametrów.

Tabela 7.5 przedstawia wyniki obliczenia średniej liczby referencji I_{21} w infoboksach na określone tematy w poszczególnych wersjach językowych Wikipedii.

W zależności od tematu i wersji językowej Wikipedii liczba odniesień jest różna. W niektórych tematach określone wersje językowe praktycznie nie używają referencji w infoboksach. Na przykład w angielskiej Wikipedii tylko ok 18 511 z 137 972 infoboksów posiada co najmniej

Tabela 7.6. Średnia liczba unikatowych referencji I_{22} w infoboksach na różne tematy w poszczególnych wersjach językowych Wikipedii.

Temat	BE	EN	PL	RU	UK
Albumy	.22	.119	.952	.567	.176
Firmy	.273	.335	.247	.304	.248
Filmy	.398	.329	.131	.135	.284
Uniwersytety	.194	.526	.29	.23	.255
Gry wideo	1.66	.54	.876	.656	.526

Źródło: Obliczenia własne.

jedną referencję. W rezultacie średnia liczba referencji I_{21} w tych infoboksach o albumach lub grach wideo w największej wersji językowej (angielskiej) może być w 2-3 razy mniejsza niż w innych rozpatrywanych wersjach językowych Wikipedii.

Innym interesującym przykładem jest białoruska i polska Wikipedia z infoboksami opisującymi gry wideo. Biorąc pod uwagę średnią wartość I_{21} prawie wszystkie z tych infobosków muszą mieć co najmniej 2 referencje. Jednak stosunkowo wysoka średnia I_{21} związana z pewną częścią infoboksów, które mają dużą liczbę przypisów. W polskiej wersji około 10% infoboksów na temat gier wideo ma ponad 6 referencji. Istnieje nawet infoboks w tej wersji językowej z prawie 40 referencjami (z artykułu „StarCraft II: Wings of Liberty"). W białoruskiej Wikipedii 3 z 50 infoboksów z grami wideo ma ponad 10 referencji.

Tabela 7.6 pokazuje wyniki obliczenia średniej liczby unikatowych referencji w poszczególnych tematach w różnych wersjach językowych Wikipedii.

W porównaniu z wynikami obliczeń średnich I_{21}, tabela 7.6 pokazuje niższe wartości. Ta różnica wynika z faktu, że czasami dwa lub więcej parametrów określonego infoboksu mogą mieć wspólne referencje jako źródło. Różnice między tabelami 7.5 i 7.6 pokazują również, jak często społeczność Wikipedii używa wspólnych źródeł do opisywania różnych parametrów konkretnego infoboksu w każdym języku. Na przykład w polskiej Wikipedii infoboks dotyczący gier wideo w średnio jedno źródło może wystąpić jako 2 referencji w konkretnym infoboksie. Istnieją jednak również takie przypadki, w których wszystkie lub prawie wszystkie referencje w ramach konkretnego infoboksu są unikalne. Dotyczy to m.in. infoboksów o albumach oraz uniwersytetach w języku białoruskim, infoboksach o filmach w języku rosyjskim.

Biorąc pod uwagę miarę kompletności I_1 z poprzedniej subsekcji, została obliczona średnia wartość liczby referencji na parametr I_{23}. Tabela 7.7 przestawia wyniki takich obliczeń.

Tabela 7.7. Średnia liczba referencji na parametr I_{23} w infoboksach na różne tematy w poszczególnych wersjach językowych Wikipedii.

Temat	BE	EN	PL	RU	UK
Albumy	.039	.015	.098	.054	.019
Firmy	.076	.106	.051	.121	.249
Filmy	.04	.041	.035	.014	.034
Uniwersytety	.03	.095	.04	.036	.054
Gry wideo	.402	.103	.214	.218	.159

Źródło: Obliczenia własne.

Wyniki pokazują, że relatywnie częściej użytkownicy Wikipedii wprowadzają referencje do parametrów infoboksów dotyczących gier wideo (zwłaszcza w języku białoruskim, polskim i rosyjskim) oraz firm (szczególnie w języku ukraińskim).

Niżej opisane są inne sposoby mierzenia wiarygodności infoboksów, które będą używane w rozprawie:

- I_{24} - całkowita liczba referencji dzielona przez liczbę uzupełnionych parametrów infoboksu z uwzględnieniem wag: $I_{24} = I_{21}/I_2$,
- I_{25} - długość w bajtach kodu źródłowego, opisującego referencje w infoboksie.
- I_{26} - długość według liczby znaków kodu źródłowego, opisującego referencje w infoboksie.

7.3.3 Aktualność

Miary związane z aktualnością mogą pomóc określić, na ile dane umieszczone w infoboksach są zbieżne z rzeczywistym stanem. Dotyczy to przede wszystkim danych, które zmieniają się w czasie. Przykładem może być liczba ludności lub prezydent miasta. Im więcej takich parametrów zawiera infoboks, tym częściej będzie on edytowany przez użytkowników.

Aktualność infoboksu można określić na podstawie analizy ostatnich zmian, które dotyczyły jego danych. Wikipedia pozwala na odczytanie wszystkich wersji historycznych artykułów, co pozwala na przeanalizowanie również zmian, które dotyczyły poszczególnych parametrów infoboksu. W związku z tym do wymiaru związanego z aktualnością infoboksu, będziemy wliczać następujące miary:

- I_{31} - liczba edycji w ciągu ostatnich 30 dni,

- I_{32} - liczba edycji w ciągu ostatnich 90 dni,
- I_{33} - liczba edycji w ciągu ostatnich 180 dni,
- I_{34} - liczba edycji w ciągu ostatnich 365 dni,
- I_{35} - liczba edycji w ciągu ostatnich 5 lat,
- I_{36} - czas ostatniej zmiany w ramach infoboksu.

7.3.4 Relewancja

Podobnie jak i w przypadku artykułów można badać, na ile relewantny czy popularny jest infoboks z punktu widzenia użytkowników. Jednak w odróżnieniu od artykułów liczba miar w tym wymiarze jest znacznie mniejsza i dotyczy głównie autorów, którzy wprowadzali zmiany do infoboksu. Na podstawie historii edycji, można określić nie tylko różnice pomiędzy poszczególnymi wersjami, ale również dowiedzieć się o autorach tych zmian.

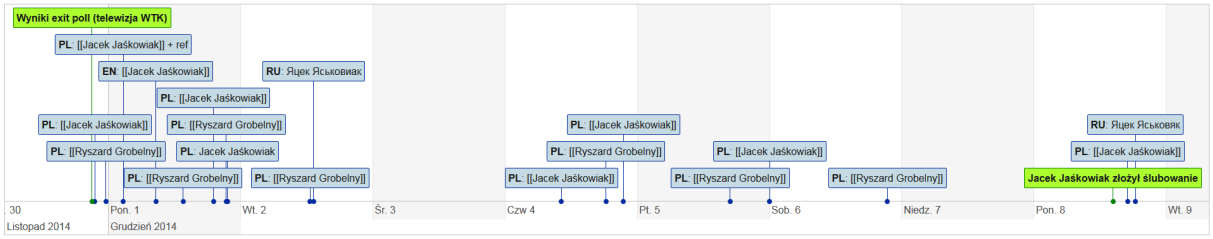
W celu mierzenia relewantności infoboksu, zostały obliczone następujące miary:

- I_{41} - liczba unikatowych autorów dokonujących zmian w ciągu ostatnich 30 dni,
- I_{42} - liczba unikatowych autorów dokonujących zmian w ciągu ostatnich 90 dni,
- I_{43} - liczba unikatowych autorów dokonujących zmian w ciągu ostatnich 180 dni,
- I_{44} - liczba unikatowych autorów dokonujących zmian w ciągu ostatnich 365 dni,
- I_{45} - liczba unikatowych autorów dokonujących zmian w ciągu ostatnich 5 lat,

7.4 Analiza jakości poszczególnych parametrów infoboksów

Każdy parametr infoboksu również może być niezależnie oceniany pod kątem jakości. Jednym z ważnych wymiarów jakości jest aktualność. Może być ona mierzona na podstawie wartości określonych parametrów. Na przykład często w infoboksach, które opisują miasta, znajduje się parametr, który wskazuje na datę (rok), kiedy została przeprowadzone liczenie ludności (np. spis powszechny). Jednak większość parametrów nie ma takiej dodatkowej informacji.

W tych samych infoboksach o miastach nie ma podanej wprost informacji, kiedy została wprowadzona wartość parametru wskazująca na prezydenta miasta. To może być szczególnie istotne w okresach wyborów samorządowych, kiedy zostały ogłoszone wyniki wyborów, ale oficjalnie nowa osoba jeszcze nie pełni tej roli. Wykres 7.3 pokazuje historię zmian parametru „zarządzający” w infoboksie o Poznaniu w rozpatrywanych wersjach językowych Wikipedii od momentu ogłoszenia wyników exit pool w telewizji WTK do złożenia ślubowania przez nowego



Rysunek 7.3. Historia zmian parametru „zarządzający” infoboksu o Poznaniu w wybranych wersjach językowych Wikipedii od momentu ogłoszenia wyników exit pool w telewizji WTK do złożenia ślubowania przez nowego prezydenta Poznania.

Źródło: Opracowanie własne na podstawie danych historycznych Wikipedii.

prezydenta Poznania. Na podstawie tego wykresu można zauważyć, że w polskiej wersji najszybciej zostały wprowadzone zmiany w infoboksie po umieszczeniu wiadomości na portalach internetowych. Dodatkowo można zauważyć, że w polskojęzycznej Wikipedii w przedstawionym okresie nie było zgodności co do wartości parametru „zarządzający” w infoboksie w artykule o Poznaniu, ponieważ formalnie został wybrany nowy prezydent miasta, jednak pełnić swoją funkcji może dopiero po złożeniu ślubowania. W angielskiej Wikipedii nie było kontrowersji na ten temat i nazwisko nowego prezydenta zostało wpisane po ogłoszeniu wyników wyborów, jednak nieco później niż to zrobiła polskojęzyczna wersja. Rosyjska Wikipedia w badanym okresie dwa razy zmieniła wartość parametru o zarządzającym miastem. Pierwsza wynikała z ogłoszenia wyników głosowania, a druga zmiana wynikała z drobnej poprawki nazwiska, zgodnie z zasadami transliteracji do cyrylicy. Co do białoruskiej oraz ukraińskiej Wikipedii - nie zanotowano tam zmian po wyborach, a nowe wartości ukazały się znacznie później. Jest to związane z tym, że wpisywanie wartości tego parametru w białoruskiej i ukraińskiej wersji infoboksu nie jest obowiązkowe – wartość ta może być bowiem automatycznie wstawiana z Wikidanych, gdzie wartość była zaktualizowana prawie po 3 latach od momentu ogłoszenia wyników wyborów.

Aktualność może być mierzona nie tylko na podstawie ostatniej zmiany konkretnego parametru, ale również analizą tzw. zmienności tego parametru w czasie. Tabela 7.8 pokazuje najczęściej zmieniane parametry infoboksów w artykułach o polskich miastach w angielskiej oraz polskiej Wikipedii w ciągu ostatnich 5 lat. Z tabeli wynika, że wśród najbardziej zmiennych parametrów są te, które są związane z populacją (populacja, liczba ludności, rok populacji). Ciekawym jest również przypadek parametru „województwo”, który jest drugim najczęściej zmienianym parametrem w ramach danego infoboksu w polskojęzycznej Wikipedii. Związano jest to w większości przypadków ze zmianą stylu zapisu wartości w tym parametrze: pierwotnie wsta-

wiano link do artykułu Wikipedii opisującego konkretne województwo, ale później użytkownicy zdecydowali zostawić tylko wartość tekstową w infoboksach danego typu.

Tabela 7.8. Najczęściej zmieniane parametry infoboksów w artykułach o polskich miastach w angielskiej oraz polskiej Wikipedii w ciągu ostatnich 5 lat.

Polska Wikipedia	
Nazwa parametru	**Liczba zmian**
populacja	2217
województwo	1020
rok (dla populacji)	1004
gęstość	687
gmina	632
zarządzający	388
powierzchnia	329
opis zdjęcia	233
zdjęcie	202
aglomeracja	201
liczba ludności	196
nazwa	187
prawa miejskie	161

Angielska Wikipedia	
Nazwa parametru	**Liczba zmian**
image_skyline	276
image_caption	273
population_total	256
leader_name	190
population_as_of	185
name	174
subdivision_name1	70
website	64
image_shield	53
motto	52
population_metro	44
area_total_km2	37
postal_code	33

Źródło: Opracowanie własne.

Analiza zmienności infoboksów w czasie także może pomóc w wykrywaniu nazw parametrów, które stosowane były wcześniej. Na przykład parametr o nazwie „populacja” ma największą zmienność w czasie, jednak obecnie ma nazwę „liczba ludności”. Warto zaznaczyć, że przy wstawieniu starej nazwy do infoboksu, wartość nie pojawia się dla czytelników, co oznacza, że stosunkowo niedawno zostały wprowadzone zasadnicze zmiany do struktury infoboksu tego typu.

Analizy zmienności mogą również pokazać, w której wersji językowej użytkownicy lepiej dbają o aktualność danych na określony temat. Z tabeli 7.8 wynika, że w polskiej wersji Wikipedii w artykułach o polskich miastach dane są aktualizowane częściej niż w największe angielskiej wersji językowej tej encyklopedii.

7.5 Podsumowanie

W niniejszym rozdziale zostały omówione miary jakości infoboksów dotyczące kompletności, wiarygodności, aktualności, relewancji. Wymiary jakości infoboksów zostały dobrane w taki

sposób, aby istniała możliwość porównywania ich z odpowiednimi wymiarami jakości artykułów.

Ekstrakcja tych miar jest możliwa przy wykorzystaniu specjalnych skryptów, przygotowanych dla potrzeb niniejszej rozprawy. Źródło tych miar są kopie zapasowe Wikipedii (dump), które zawierają m.in. kody źródłowe artykułów z infoboksami i historię edycji artykułów. Większość z tych miar jest oryginalną propozycją autora niniejszej rozprawy i nie były one wykorzystane w innych podobnych badaniach.

Badanie jakości infoboksu może również się odbywać na poziomie analizy poszczególnych parametrów. W tym przypadku bardzo przydatnym może być mierzenie aktualności oraz zmienności poszczególnych parametrów infoboksu.

Następny rozdział poświęcony jest budowaniu modeli jakości infoboksów na podstawie różnym miar. Zostanie również poruszona kwestia współzależności pomiędzy niektórymi miarami jakości artykułów oraz infoboksów.

Rozdział 8

Budowanie modeli jakości infoboksów

W niniejszym rozdziale zostaną przedstawione zbudowane modele jakości infoboksów na podstawie wcześniej opisanych miar.

Tworzenie modeli było poprzedzone analizą jakości infoboksów w różnych językach oraz tematach.

8.1 Wprowadzenie

W celu dalszych analizy zostały wybrane infoboksy z artykułów jednego tematu, które spełniają łącznie następujące warunki:

- artykuły posiadają infoboks,
- artykuły są opisane we wszystkich rozpatrywanych wersjach językowych,
- infoboksy muszą być tego samego rodzaju w ramach danej wersji językowej.

Ostatni z wyżej opisanych kryteriów mówi o tym, że artykuły muszą mieć infoboks tego samego rodzaju. Istnieją też infoboksy podobnego rodzaju, np. infoboksy opisujące ludzi: politycy, naukowcy, artyści i inne. Takie infoboksy mogą posiadać wspólne parametry (takie jak np. data i miejsce urodzenia), jednak można znaleźć dodatkowe unikatowe parametry dla każdego z tych infoboksów.

W celu dalszej analizy zostały wybrane artykuły na 6 tematów: miasta Polski, miasta Ukrainy, miasta Rosji, firmy, uniwersytety, gry komputerowe.

Na przykład w celu utworzenia zbioru artykułów o miastach Polski były dobrane artykuły z polskojęzycznej Wikipedii, które posiadają infoboks o nazwie „Polskie miasto infobox”. Całkowita liczba takich artykułów to 934. Warto zaznaczyć, że w polskojęzycznej Wikipedii istnieje

również infoboks o nazwie „Wieś infobox”, który również dotyczy polskich miejscowości, które jednak są zazwyczaj mniejsze niż w przypadku artykułów z infoboksem „Polskie miasto infobox”. Jeżeli zostawić artykuły, które mają co najmniej 5 rozpatrywanych wersji językowych, to zostanie 200 artykułów.

Białoruska Wikipedia jest mniej rozwiniętą w porównaniu do 4 pozostałych rozpatrywanych wersji językowych tej encyklopedii. Jeżeli pominiemy wymóg posiadania przez artykuły o polskich miastach wersji białoruskiej, to liczba artykułów będzie znacznie większa - 901.

W dalszych opisach, będą używane następujące skróty:

- **ARPU** - artykuły, posiadające co najmniej 4 wersje językowe: angielska, rosyjska, polska, ukraińska,
- **ARPUB** - artykuły, posiadające co najmniej 5 wersji językowych: angielska, rosyjska, polska, ukraińska, białoruska

Tabela 8.1 przedstawia liczbę artykułów Wikipedii na określony temat w wersji ARPU oraz ARPUB.

Tabela 8.1. Liczba artykułów Wikipedii na określony temat w wersji ARPU oraz ARPUB (4 oraz 5 wybranych wersji językowych odpowiednio).

Temat	ARPU	ARPUB
Miasta Polski	901	200
Miasta Rosji	885	798
Miasta Ukrainy	1304	303
Firmy	1387	125
Uniwersytety	593	201
Gry komputerowe	1025	27
Razem	**6095**	**1654**

Źródło: Obliczenia własne.

Dla każdego zbioru danych będą budowane modele jakości na podstawie miar infoboksów oraz oddzielnie na podstawie miar infoboksów wspólnie z miarami artykułów. Będą zatem rozpatrywane dwie wersje modeli:

- **podstawowa** - zawiera tylko miary infoboksów (I),
- **rozszerzona** - zawiera miary infoboksów (I_x) oraz miary artykułów (A_x).

Każda z tych wersji jest rozpatrywana oraz opisywana oddzielnie w następnych podsekcjach.

W celu zbudowania modeli jakości infoboksów zostało losowo wybrane po 200 artykułów z każdego rozpatrywanego tematu w wersji ARPU - razem to 1200 artykułów.

Ręczna adnotacja jakości polegała na wyborze najlepszej wersji językowej. Były brane pod uwagę 4 wymiary jakości: kompletność, aktualność, wiarygodność, relewancja. Zazwyczaj polskie miasta są najlepiej opisane w polskojęzycznej Wikipedii we wszystkich 4 wymiarach. Jednak są przykłady, kiedy inne wersje językowe mają podobną jakość w ramach niektórych wymiarów.

Na przykład aktualność danych w momence przeprowadzania analizy była podobna pomiędzy wersjami angielską i polską w infoboksach następujących miast: Warszawa, Łódź, Rzeszów. Podobną aktualność danych można było też spotkać pomiędzy polską a rosyjską oraz ukraińską wersją w infoboksach opisujących Wrocław, Gdańsk, Lublin i Opole. W takich przypadkach brana była pod uwagę historia zmian tych danych. W większości przypadków polska wersja wcześniej posiadała zaktualizowane dane (w następnej sekcji osobno rozpatrywana jest kwestia analizy zmienności poszczególnych parametrów infoboksu). Kolejny przykład dotyczy wiarygodności – infoboksy o takich miastach, jak Sandomierz, Kostrzyn nad Odrą, Środa Wielkopolska, Tarnobrzeg, zawierały największą liczbę referencji wśród innych niż polska wersjach językowych. Nawet jeżeli w jednym wymiarze jakości (np. aktualność) wygrywała inna wersja niż polska, to w pozostałych zazwyczaj ustępowała (np. kompletność, wiarygodność). W takich przypadkach najlepszej wersją była odznaczona ta, która była najlepszą według większości wymiarów jakości oraz miała bardziej poprawne dane.

Zmienna zależna została skonstruowana jako zmienna dychotomiczna, dla każdej wersji językowej, i może zawierać jedną z dwóch wartości: „Lepsza" albo „Gorsza". Te wartości były nadane ręczne i w ramach jednego artykułu tylko jedna wersja językowa może mieć kategorię „Lepsza", reszcie przypisano kategorię „Gorsza". Zmiennymi niezależnymi były miary infoboksów lub w przypadku wariantu rozszerzonego – również miary jakości artykułów.

Jakość infoboksów była porównywana do jakości artykułów, które je zawierały. Wcześniej, w rozdziale nr 4 „Metody określenia jakości artykułów Wikipedii" było pokazane, że duża liczba artykułów w rozpatrywanych wersjach językowych nie posiada ocen jakości. W związku z tym, wszystkie analizowane artykuły z infoboksami były ocenione przy pomocy zbudowanych w ramach niniejszej pracy modeli jakości artykułów (J1-J5). Dodatkowo były tez brane pod uwagę wszystkie mary jakości artykułów (A1-A138), w celu zidentyfikowania najbardziej istotnych z punktu widzenia oceny jakości infoboksu.

Dla każdego artykułu zostało wyekstrahowane ponad 100 miar dotyczących jakości wybranych artykułów (A1-A138) oraz jakości infoboksów zawartych w tych artykułach (I1-I45). Wszystkie te miary występują w modelach jako zmienne niezależne. W celu uzyskania zmiennej zależnej, której jest jakość infoboksu, zostały ręcznie odznaczone te wersje językowe, które posiadały dane o najwyższej jakości. Przy budowaniu modeli został wykorzystany algorytm Lasu Losowego (Random Forest).

Przed zbudowaniem modeli została przeprowadzona normalizacja min-max każdej miary. Przeprowadzono liniową transformację pierwotnych danych do przedziału $[0, 1]$, która bierze pod uwagę minimalną oraz maksymalną wartość miary w każdej z rozpatrywanych wersji językowych poszczególnych artykułów oraz infoboksów. Na przykład, jeżeli rozpatrujemy 4 wersje językowe artykułu „Poznań" i bierzemy pod uwagę *długość artykułu w bajtach*, to absolutne wartości oraz wartości po normalizacji min-max (w nawiasach) tej miary w poszczególnych wersjach językowych wyglądają następująco:

- wersja angielska: 76.436 (0.66)
- wersja polska: 183.902 (1)
- wersja rosyjska: 16.746 (0)
- wersja ukraińska: 32.056 (0.45)

Ocena jakości będzie modelowana jako prawdopodobieństwo przynależności do jednej z dwóch kategorii „Lepsza"„ lub „Gorsza".

8.2 Wersja podstawowa modelu

W wersji podstawowej model jakości był budowany na podstawie tylko miar jakości infoboksów o miastach Polski, Ukrainy, Rosji, firmach, grach wideo oraz uniwersytetach dla 4 wersji językowych w ramach zbioru danych ARPU.

Każdy z badanych artykułów posiadał wyekstrahowane miary jakości artykułów oraz infoboksów. Miarą zależną była ocena w skali dychotomicznej, która wskazywała na najlepszą wersję językową dla każdego artykułu. Zbiór danych wynosił 4800 rekordów (4 wersje językowe na każdy z 1200 artykułów, 200 artykułów z każdego z rozpatrywanych tematów).

Precyzja modelu jakości infoboksów przy użyciu algorytmu lasa losowego wyniosła 93,1%. W tabeli 8.2 przedstawiono 10 najważniejszych miar z podaniem ważności każdego z nich w skali od 0 do 100.

Tabela 8.2. Najważniejsze miary w modelu jakości infoboksów w wersji podstawowej dla 4 wersji językowych w ramach zbioru danych ARPU.

Skrót	Opis miary	Ważność
I_4	Mediana liczby wszystkich wypełnionych parametrów I_1 dla danego typu infoboksu w wybranej wersji językowej Wikipedii	100
I_2	Stosunek sumy liczby wypełnionych parametrów z wagami do liczby wszystkich parametrów w ramach infoboksu	96,43
I_1	Liczba wypełnionych parametrów infoboksu	77,35
I_3	Liczba wszystkich wypełnionych parametrów IN_1, z wyjątkiem tych, które mają wagę mniejszą niż 0.1	73,19
I_{11}	Liczba wszystkich wypełnionych parametrów z uwzględnieniem wag (I_2), z wyjątkiem tych parametrów, które w co najmniej 50% przypadkach mają podobne wartości	67,25
I_6	Średnia długość wartości infoboksu: $I_6 = I_5/I_1$	59,43
I_{36}	Czas ostatniej zmiany w ramach infoboksu	59,21
I_7	Liczba linków do artykułów Wikipedii w wartościach infoboksu	56,51
I_9	Liczba szablonów w wartościach parametrów	56,33
I_5	Długość w bajtach wszystkich wartości parametrów infoboksu	55,58

Źródło: Obliczenia własne przy użyciu WEKA.

Powyższe dane wskazują, że jakość infoboksu jest głównie określana przez miary dotyczące kompletności infoboksu.

W dalszej kolejności zostały zbudowane modele jakości tylko na podstawie miar z określonego wymiaru jakości: kompletność (miary I1-I20), wiarygodność (miary I21-I26), aktualność (miary I31-I36), relewancja (miary I41-I45). Model z miarami relewancji miał najwyższą precyzję spośród innych wymiarów, jednak według wskaźników ROC Area oraz PRC Area modele z miarami kompletności mają przewagę nad innymi wymiarami. Tabela 8.3 pokazuje wskaźniki jakości modeli przy wykorzystaniu miar z poszczególnych wymiarów oraz najważniejsze miary w danych modelach.

Warto zaznaczyć, że miary kompletności uzyskały w modelu ogólnym wysokie pozycje w rankingu najważniejszych miar (tab. 8.2), ale do uzyskania najwyższej precyzji, należy stosować również miary z innych wymiarów.

Na podstawie zbudowanego modelu jakości w wersji podstawowej zostały ocenione wszystkie artykuły w wersji ARPU, które były opisane wcześniej (patrz tabela 8.1). Ogólne statystyki z liczebnością najlepszych wersji językowych określonych przez model jakości infoboksów w wersji podstawowej pokazuje tabela 8.4. W rozdziale nr 11 („Ewaluacja metod") zostanie prze-

Tabela 8.3. Wskaźniki jakości modelu jakości infoboksów w wersji podstawowej oraz najważniejsze miary (NM) przy wykorzystaniu miar z określonych wymiarów jakości.

Wymiar	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	NM
Kompletność	0,896	0,897	0,896	0,722	0,949	0,951	I_4
Wiarygodność	0,796	0,804	0,798	0,453	0,734	0,774	I_{23}
Aktualność	0,900	0,902	0,901	0,733	0,930	0,928	I_{36}
Relewancja	0,912	0,910	0,911	0,766	0,920	0,915	I_{41}
Wszystkie	**0,932**	**0,932**	**0,932**	**0,818**	**0,979**	**0,98**	I_4

Źródło: Obliczenia własne w programie WEKA.

prowadzona ewaluacja metody, w której zostaną porównane wyniki otrzymane przez model oraz wyniki otrzymane od ekspertów.

Tabela 8.4. Liczebność ocenionych przez model najlepszych wersji językowych określonych przez model jakości infoboksów w wersji podstawowej.

Temat	Najlepsze wersje językowe			
	EN	PL	RU	UK
Miasta Polski	0	901	0	0
Miasta Ukrainy	57	14	308	925
Miasta Rosji	71	0	811	3
Firmy	1157	42	172	16
Uniwersytety	453	44	76	20
Gry komputerowe	889	38	82	16
Razem	**2627**	**1039**	**1449**	**980**

Źródło: Oobliczenia własne przy użyciu WEKA.

8.3 Wersja rozszerzona modelu

W wersji rozszerzonej modelu oprócz miar jakości infoboksów wykorzystano również dotyczące artykułów (A1-A133). Precyzja takiego modelu z 4 wersjami językowymi o polskich miastach nieco wzrosła i wyniosła 96,2%. Tabela 8.5 przedstawia 10 najistotniejszych miar w tym modelu.

Warto zwrócić uwagę na to, że na listę najważniejszych miar trafiły głównie miary jakości artykułów. W szczególności, wśród najważniejszych znalazły się miary związane z popytem na informację (m.in to miary A84, A85, A87).

Tabela 8.5. 10 najważniejszych miar w modelu jakości infoboksów w wersji rozszerzonej dla 4 wersji językowych w ramach zbioru danych ARPU.

Skrót	Opis miary	Ważność
A84	Suma odwiedzin za ostatni rok	100,00
A85	Mediana odwiedzin w ciągu ostatnich 90 dni	89,51
A87	Mediana odwiedzin w ciągu ostatnich 365 dni bez dni z brakiem odwiedzin	68,91
A138	Gęstość referencji artykułu znormalizowana do wskaźnika syntetycznego	64,04
A86	Mediana odwiedzin w ciągu ostatnich 365 dni	62,17
A136	Liczba obrazków artykułu znormalizowana do wskaźnika syntetycznego	59,93
A27	Długość abstraktu według liczby znaków	59,93
I10	Liczba wszystkich wypełnionych parametrów (I_1), z wyjątkiem tych, które w 50% przypadkach mają podobne wartości	59,93
A26	Długość abstraktu w bajtach	59,55
A77	Liczba unikatowych autorów w ciągu ostatnich 180 dni	59,55

Źródło: Obliczenia własne przy użyciu WEKA

Jeżeli chodzi o miary J1-J5, które zawierają obliczone oceny jakości według różnych standardów, to największą ważność uzyskał wskaźnik syntetyczny (J5) - 53,6. Mniejszą ważność posiadają oceny według standardów rosyjskiej oraz angielskiej Wikipedii przy zmiennej nominalnej - 21,7 oraz 18,4 odpowiednio. Jeżeli jakość artykułów rozpatrywać jako zmienną dychotomiczną, do dla wersji rosyjskiej oraz wersji polskiej ważność miary wynosi odpowiednio 12,7 oraz 2,6. To pokazuje, że **system ocen artykułów stosowanych w ramach różnych wersji językowych Wikipedii nie jest istotny dla oceny jakości infoboksów**.

Podobnie jak i dla wersji podstawowej, przy użyciu modelu jakości infoboksu w wersji rozszerzonej zostały określone najlepsze wersje językowe w ramach każdego z rozpatrywanych tematów. Wyniki ocen przy użyciu modelu rozszerzonego pokazuje tabela 8.6.

Porównując tabele 8.4 oraz 8.6 można zauważyć różnice. Np. w przypadku miast Polski, wersja rozszerzona dla 2 artykułów określiła wersje angielskie jako najlepsze, wówczas gdy wersja podstawowa nie wykryła innych lepszych wersji językowych oprócz polskiej w ramach danego tematu.

W rozdziale nr 11 („Ewaluacja metod") zostaną porównywane model jakości infoboksów w wersji podstawowej oraz model jakości infoboksów w wersji rozszerzonej z wynikami ocen ekspertów.

Tabela 8.6. Liczebność ocenionych przez model najlepszych wersji językowych określonych przez model jakości infoboksów w wersji rozszerzonej.

Temat	Najlepsze wersje językowe			
	EN	PL	RU	UK
Miasta Polski	2	899	0	0
Miasta Ukrainy	22	14	355	913
Miasta Rosji	21	0	864	0
Firmy	1235	22	122	8
Uniwersytety	471	34	64	24
Gry komputerowe	910	8	107	0
Razem	**2661**	**977**	**1512**	**945**

Źródło: Obliczenia własne przy użyciu WEKA.

8.4 Współzależność miar jakości infoboksów i artykułów

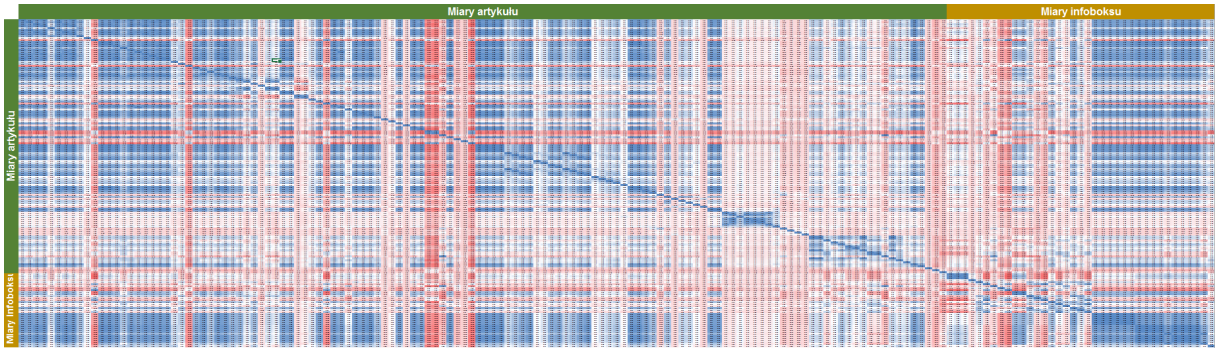
W tej sekcji przeprowadzono analizę współzależności miar jakości infoboksów z miarami jakości artykułów.

Do pokazania współzależności pomiędzy dwoma zmiennymi (w naszym przypadku - miarami) można obliczać współczynnik korelacji. Jeżeli współczynnik jest mniejszy niż 0, to oznacza odwrotną korelację pomiędzy zmiennymi - zwiększenie jednej miary musi powodować zmniejszenie drugiej. Siła współzależności pomiędzy miarami jest największa przy wartości współczynnika równej 1 (przy pozytywnej korelacji) lub -1 (przy ujemnej korelacji). Rozróżniamy następujące poziomy współzależności w zależności od wartości współczynnika korelacji (Jackson, 2014):

- $,70 - 1,00$ - silna współzależność,
- $0,30 - 0,69$ - umiarkowana współzależność,
- $0,0 - 0,29$ - słaba lub brak współzależności,

Na podstawie wcześniej przygotowanego zbioru danych, który był opisany w sekcji nr 8.1., została zbudowana macierz korelacji. Ta macierz zawiera 143 miary jakości dotyczące artykułów oraz 37 miar dotyczące infoboksów. Rysunek 8.1 pokazuje ogólną macierz korelacji innych miar jakości infoboksów i artykułów.

Tabela 8.7 pokazuje korelację pomiędzy wybranymi miarami infoboksów oraz wybranymi miarami jakości artykułów. W celu oszczędzania miejsca zostały wybrane tylko najważniej-



Rysunek 8.1. Macierz korelacji miar jakości artykułów i infoboksów w ramach zbioru danych ARPU

Źródło: Opracowanie własne.

sze miary infoboksów z poszczególnych wymiarów jakości (kompletność, wiarygodność, aktualność, relewancja), które zostały zidentyfikowane przy budowaniu modeli jakości w wersji podstawowej (por. tabelę 8.3). Dodatkowo na liście miar jakości artykułów zostawiono tylko te, które miały silną współzależność z wybranymi miarami infoboksów na co najmniej przyzwoitym poziomie (ponad 0,3).

Tabela 8.7. Korelacja pomiędzy wybranymi miarami infoboksów oraz wybranymi miarami jakości artykułów. Skróty w nawiasach: K - kompletność, W - wiarygodność, A - aktualność, R - relewancja.

Miara	I_4 (K)	I_{23} (W)	I_{36} (A)	I_{41} (R)
A_{36} - Linki przychodzące z przestrzeni nazw ns0	0,049	0,338	0,298	0,715
A_{54} - Linki przychodzące ze wszystkich rozpatrywanych przestrzeń nazw	0,045	0,339	0,300	0,732
A_{55} - Linki przychodzące z artykułów Wikipedii	0,049	0,313	0,384	0,772
A_{63} - Liczba edycji artykułu	0,041	0,338	0,376	0,809
A_{64} - Liczba drobnych edycji	0,038	0,363	0,386	0,783
A_{73} - Liczba unikatowych autorów	0,029	0,337	0,385	0,814
A_{74} - Liczba unikatowych autorów anonimowych	0,022	0,311	0,368	0,785
A_{84} - Suma odwiedzin w ciągu ostatnich rok	0,043	0,300	0,337	0,732
A_{85} - Mediana odwiedzin w ciągu ostatnich 90 dni	0,043	0,295	0,332	0,720
A_{86} - Mediana odwiedzin w ciągu ostatnich 365 dni	0,038	0,290	0,332	0,708
A_{87} - Mediana odwiedzin w ciągu ostatnich 365 dni bez dni z brakiem odwiedzin	0,039	0,297	0,340	0,715
A_{89} - Liczba wszystkich referencji w abstrakcie	0,113	0,704	0,242	0,363

Źródło: Obliczenia własne.

Analizy pokazują, że niektóre z wybranych miar jakości infoboksów są mocno skorelowane z miarami jakości artykułów, które określają popyt (liczba odwiedzin), aktualność (liczba edycji), relewancja (liczba autorów), wiarygodność (liczba referencji).

Warto zaznaczyć, że spośród ocen jakości artykułów ($J1 - J5$), największą wartość współczynnika korelacji posiada wskaźnik syntetyczny ($J5$). Tabela 8.8 przedstawia wyniki obliczenia

współczynników korelacji pomiędzy wybranymi miarami infoboksów oraz ocenami jakości artykułów według różnych modeli.

Tabela 8.8. Korelacja pomiędzy wybranymi miarami infoboksów oraz ocenami jakości artykułów według różnych modeli. Skróty w nawiasach: K - kompletność, W - wiarygodność, A - aktualność, R - relewancja.

Miara	I_4 (K)	I_{23} (W)	I_{36} (A)	I_{41} (R)
J_1 - Jakość artykułu według modelu angielskiej Wikipedii z kategorialnej zmiennej zależnej w skali od 1 do 7	-0,009	0,242	0,243	0,469
J_2 - Jakość artykułu według modelu angielskiej Wikipedii z dychotomicznej zmiennej zależnej w skali od 0 do 1	-0,012	0,085	0,095	0,228
J_3 - Jakość artykułu według modelu rosyjskiej Wikipedii z kategorialnej zmiennej zależnej w skali od 1 do 7	0,007	-0,091	-0,087	-0,160
J_4 - Jakość artykułu według modelu rosyjskiej Wikipedii z dychotomicznej zmiennej zależnej w skali od 0 do 1	0,006	-0,039	-0,027	-0,057
J_5 - Ocena jakości artykułu z wykorzystaniem miary syntetycznej	0,131	0,490	0,335	0,605

Źródło: Obliczenia własne.

W przypadku miar $J1 - J4$, które zostały zbudowane na podstawie system ocen angielskiej oraz rosyjskiej Wikipedii, mamy do czynienia ze słabą współzależnością ze wszystkimi wybranymi miarami infoboksów. Najlepsze wyniki pokazała miara jakości artykułów $J5$, obliczona na podstawie wskaźnika syntetycznego - posiada umiarkowaną korelacją z miarami infoboksów dotyczących wiarygodności, aktualności oraz relewancji.

8.5 Podsumowanie

W celu zbudowania modeli jakości infoboksów zostały użyte dwa zbiory miar: w wersji podstawowej oraz rozszerzonej. W wersji podstawowej były wybrane tylko miary infoboksów (I1-I45). Precyzja modelu wyniosła 93,1%. W wersji rozszerzonej zostały dodatkowo dodane miary artykułów (A1-A138 oraz J1-J5). W wyniku rozszerzenia zbioru uczącego precyzja modelu oceny jakości infoboksów zwiększyła się o 3,1% i wyniosła 96,2%.

Miary związane z popytem na informację okazały się najważniejszymi w otrzymanym modelu w wersji rozszerzonej. Dodatkowo modele jakości pokazały, że oceny artykułów, otrzymane od użytkowników Wikipedii, nie są istotne z punku widzenia oceny jakości infoboksów.

Eksperymenty wykazały korelację między niektórym miarami jakości artykułów oraz infoboksów. Najczęściej wymiary jakości infoboksów są mocno skorelowane z miarami pokazującymi liczbę wyświetleń artykułu, liczbę referencji (w tym z popularnych stron internetowych

oraz z używaniem specjalnych szablonów), długość tekstu artykułu, liczbę obrazków i sekcji oraz liczbę edycji wraz z liczbą unikatowych autorów.

Modele jakości zbudowane na podstawie miar infoboksów oraz artykułów o polskich miastach zostały wykorzystane do automatycznej oceny jakości infoboksów pomiędzy wersjami językowymi w innych tematach, co jest zaprezentowane w kolejnych rozdziałach.

Rozdział 9

Porównywanie informacji wielojęzycznych

W tym rozdziale zostanie opisana metoda porównywania poszczególnych wartości parametrów infoboksów pomiędzy różnymi wersjami językowymi Wikipedii.

9.1 Wprowadzenie

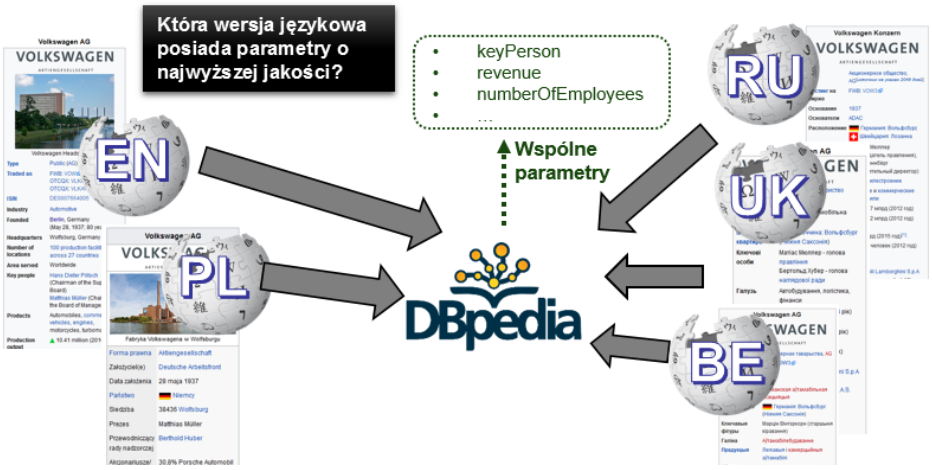
W poprzednich rozdziałach wskazano, że każda wersja językowa Wikipedia może niezależnie od innych języków definiować infoboksy wraz z nazwami parametrów, które mogą być używane w ramach szablonów. W związku z tym treści o określonym podmiocie lub wydarzeniu mogą powstawać niezależnie w każdej wersji językowej Wikipedii.

Niektóre tematy mogą być lepiej opisane w określonych wersjach językowych. Na przykład, przy budowaniu modeli jakości infoboksów (patrz rozdział nr 8 „Budowanie modeli jakości infoboksów") miasta Polski zazwyczaj są lepiej opisane w polskojęzycznej wersji Wikipedii.

Jednym z ważnych elementów opisywanej w tym rozdziale metody jest model jakości infoboksów w wersji rozszerzonej, w której jakość jest modelowana jako prawdopodobieństwo przynależności od jednej z dwóch kategorii „Lepsza" lub „Gorsza" relatywnie do innych języków. To prawdopodobieństwo p będzie wykorzystywane do obliczenia punktów jakości Q dla wersji językowej artykułu w:

- $Q(w) = p$, jeżeli artykuł został zaklasyfikowany jako „Lepszy",
- $Q(w) = 1 - p$, jeżeli artykuł został zaklasyfikowany jako „Gorszy",

Przy tym wartość prawdopodobieństwa p obliczana osobno na każdej wersji językowej każdego rozpatrzonego artykułu.



Rysunek 9.1. Ekstrakcja parametrów infoboksów opisujących firmę w różnych wersjach językowych Wikipedii oraz unifikacja do wspólnych nazw za pośrednictwem DBpedii.

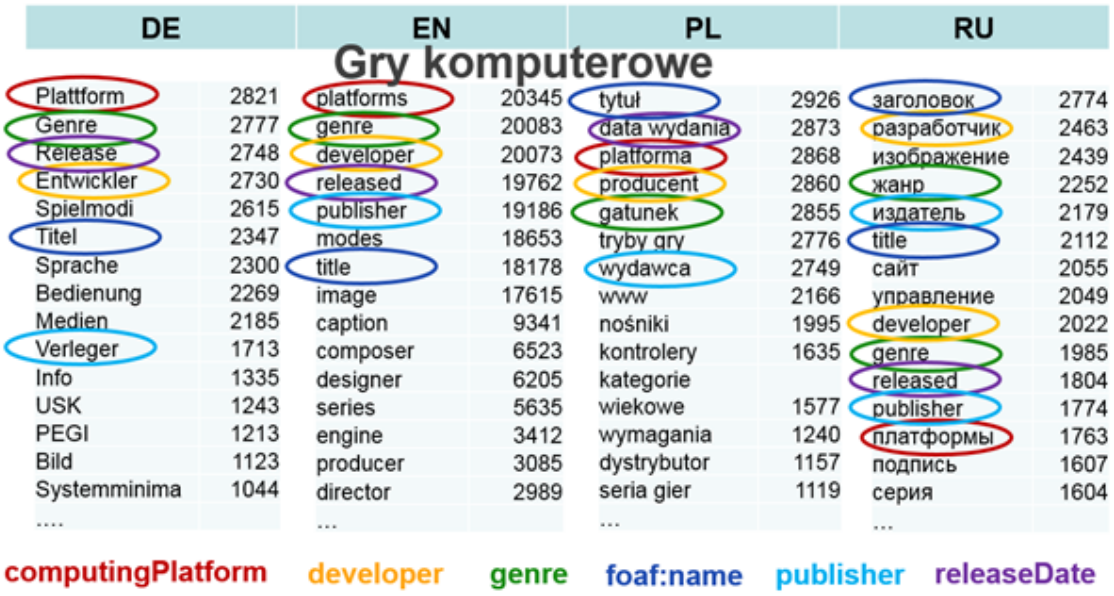
Źródło: Opracowanie własne.

9.2 Unifikacja parametrów infoboksów

Przed rozpoczęciem porównywania wielojęzycznych informacji należy zidentyfikować dane, które mogą oznaczać to samo, ale kryją się pod różnymi nazwami parametrów. Semantyczna baza DBpedia posiada mechanizm przekształcenia nazw parametrów z różnych wersji językowych do jednego ujednoliconego standardu. Ten proces nazywa się unifikacją parametrów. Rysunek 9.1 pokazuje schemat przekształcenia nazw parametrów infoboksów przy pomocy DBpedii na przykładzie 5 rozpatrywanych wersji językowych Wikipedii. Wówczas pojawia się pytanie, która wersja językowa posiada wartość konkretnego parametru o lepszej jakości od pozostałych.

Mapowania DBpedii pozwalają na unifikację nazw infoboksów, które mogą mieć różne brzmienie w zależności od wersji językowej Wikipedii. Na przykład, infoboksy o grach komputerowych najczęściej maja uzupełnione następujące parametry: tytuł, data wydania, platforma, producent, gatunek, wydawca. We wszystkich 4 rozpatrywanych wersjach językowych istnieje odpowiednia nazwa (nazwy) każdego parametru. Używając mapowań DBpedii przeprowadzimy unifikację nazw zgodnie z rys. 9.2.

Należy zaznaczyć, że w ramach jednego infoboksu w konkretnej wersji językowej mogą być używane dwie i nawet więcej nazw parametru, który wyświetla się dla czytelnika tak samo. Na przykład, w rosyjskiej Wikipedii infoboks o grach komputerowych może zawierać parametr o nazwie „title", który jest odpowiednikiem parametru „tytuł" w polskiej wersji. Dodatkowo w



Rysunek 9.2. Unifikacja nazw parametrów infoboksów o grach komputerowych w różnych wersjach językowych Wikipedii

Źródło: Opracowanie własne.

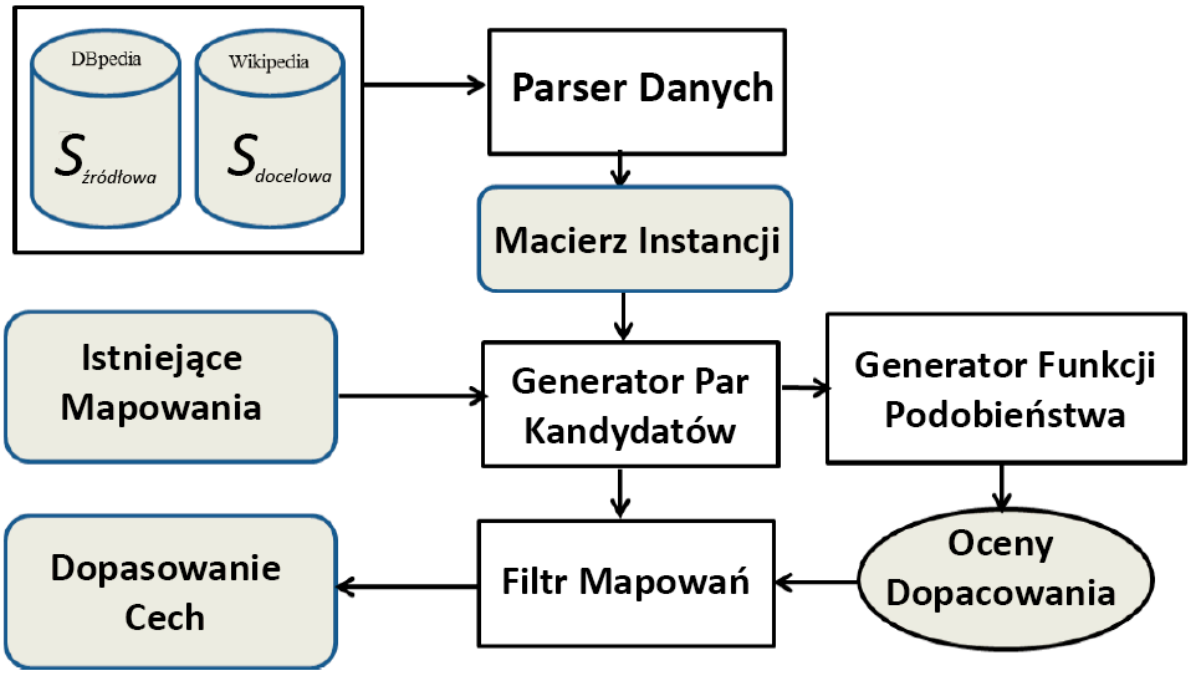
ramach tego samego infoboksu można też używać inną nazwę w cyrylicy, która w wyniku będzie tak samo wyświetlana na stronie Wikipedii.

W związku z powyższym należy zadbać o to, żeby nie tylko znaleźć odpowiedniki każdego parametru infoboksu w każdym języku, ale też zbadać inne nazwy wskazujące na ten sam parametr w ramach jednej wersji językowej.

Unifikację nazw parametrów przeprowadza się na podstawie wstępnie przygotowanych mapowań, w których są zapisane reguły przekształcenia nazw parametru określonego infoboksu oraz wersji językowej do odpowiednika, który używany jest w ramach ontologii innej bazy danych. DBpedia, jako przykład takiej bazy, na specjalnym serwisie „DBpedia Mappings”[1] proponuje użytkownikom wspólnie tworzyć reguły unifikacji nazw infoboksów w różnych językach. W sierpniu 2018 roku serwis ten umożliwiał tworzenie mapowań dla około 50 wersji językowych. Tylko dla angielskiej wersji zostały opisane reguły unifikacji parametrów dla ponad 700 infoboksów. Nie zawsze jednak takie reguły zawierają mapowania dla wszystkich parametrów w ramach określonych infoboksów. Poza tym w tym procesie mogą pojawić się różne problemy: użytkownicy mogą tworzyć złe lub niespójne mapowania, używać ontologii w sposób nieprzewidziany lub zmieniać ontologię bez uwzględnienia wszystkich możliwych konsekwencji (Paulheim, 2017).

[1] `http://mappngs.dbpedia.org`

Istnieją badania, które opisują podejścia do automatyzacji procesu tworzenia mapowań na podstawie jednej wersji językowej do wielu innych (E. Kim i Choi, 2014; N. Nguyen, Cao i Nguyen, 2018; T. Nguyen, Moreira, Nguyen, Nguyen i Freire, 2011; T.-N. Nguyen, Takeda, Nguyen, Ichise i Cao, 2016; Palmero Aprosio, Giuliano i Lavelli, 2013; Rinser, Lange i Naumann, 2013). Istnieją również badania, które pomagają zidentyfikować błędy w mapowaniach DBpedii (Rico i in., 2018). Ogólny schemat automatycznego generowania mapowań infoboksów przedstawiony został na rysunku 9.3.



Rysunek 9.3. Schemat automatycznego generowania mapowań infoboksów.

Źródło: Opracowane na podstawie (N. Nguyen, Cao i Nguyen, 2018)

Do porównywania wartości należy również zadbać o odpowiedni format wpisywanych danych. Wartości w infoboksach mogą być różnego rodzaju, m.in. to:

- odnośnik do strony zewnętrznej (np. strona WWW),
- odnośnik do strony Wikipedii (tzw. „wikilink”),
- nazwa pliku (np. obrazu),
- wartość tekstowa (imię, nazwa, cytat),
- wartość liczbowa (liczba ludności, data, dochód).

W celu porównania wartości tego samego parametru w różnych językach należy najpierw zbadać, czy wartości są porównywalne. W przypadku gdy jeden parametr ma wartość w innym

formacie (np. „13 września 2018 roku” zamiast „13.09.2018”), należy przeprowadzić standaryzację wartości zgodnie z wybranym formatem.

Jeżeli wartość parametru jest przedstawiona w postaci odnośnika do strony Wikipedii, należy sprawdzić, czy ta strona istnieje i czy posiada ona inne wersje językowe. Ponieważ w niektórych wersjach językowych nazwy własne (w tym imiona, nazwy geograficzne) często są przekształcane (czy transliterowane), taki odnośnik może pomóc w określeniu poprawnego odpowiednika wartości tego parametru w innym języku.

Jeżeli wartość parametru infoboksu ma wpisaną nazwę pliku (najczęstszej jest to obrazek), należy sprawdzić, czy ten plik istnieje. Nawet jeśli się okaże, że lepsza wersja językowa nie posiada tego obrazka, wtedy należy sprawdzić wartość odpowiedniego parametru w wersji językowej.

Dodatkowo wartości parametrów infoboksów mogą zawierać referencje czy inne elementy (szablony). W związku z tym przed porównywaniem wartości parametrów pomiędzy wersjami językowymi, należy podzielić taką wartość na główną oraz dodatkową. Ta dodatkowa może być brana pod uwagę jako dodatkowa miara jakości (np. wiarygodność, jeżeli to jest referencja), lecz obiektem porównywania będzie główna zunifikowana wartość. W zależności od wersji językowej oraz tematu infoboksu, takie dodatkowe elementy mogą mieć swój oddzielny parametr. Np. w angielskiej Wikipedii, infoboksy opisujące miasta w przeciwieństwie do polskiej wersji mają dla wskazania referencji oddzielne parametry o nazwach „population footnotes”, „population note” czy też „footnotes”.

Dodatkowymi elementami wartości mogą być nie tylko referencje. Innym przykładem takich elementów jest szablon, który może automatycznie wstawiać wartość z innego źródła (w tym. z tzw. „tabular data”), przekształcać wartość (np. konwertować z metrów na funty), wstawiać grafikę (np. flagę państwa, z którym związana jest wartość).

9.3 Metoda porównywania informacji na podstawie analizy jakości

Zunifikowane nazwy parametrów ułatwiają porównanie ich wartości w różnych językach. Jednak w przypadku rozbieżności wartości, należy wybrać tę wersję językową, która posiada najwyższą jakość. To może się odbywać na różnych poziomach analizy: jakość całego artykułu, jakość infoboksu oraz jakość poszczególnych parametrów.

Proponowana metoda porównywania informacji dla jednego artykułu zawiera następujące kroki:

- Najpierw należy wybrać artykuł oraz wersje językowe, które będą analizowane. Każdy artykuł musi posiadać infoboks podobnego rodzaju.
- Dla każdej wersji językowej artykułów należy wyekstrahować parametry z infoboksów.
- Unifikacja nazw parametrów infoboksów - nazwy musza być przekształcone do jednego wspólnego standardu.
- Unifikacja wartości parametrów infoboksów. Jeżeli wartością jest nazwa własna, wykorzystywany jest semantyczny odpowiednik tej nazwy lub stosowane jest przekształcenie tej nazwy zgodnie z zasadami transliteracji w zależności od wersji językowych.
- Dla każdego parametru mogą być dodatkowo wyekstrahowane miary jakości związane z aktualnością, kompletnością oraz weryfikowalnością.
- Przeprowadzenie porównania wartości parametrów infoboksów.
- W przypadku różnic na poziome określonych parametrów, należy wziąć pod uwagę wyniki oceny jakości wersji językowych wybranego artykułu. Preferowana jest wartość z tych wersji językowych, które uzyskały większą liczbę punktów za jakość. W celu przeprowadzenia takiej oceny należy zastosować model jakości infoboksów w wersji rozszerzonej, przy tym należy uwzględnić następujące kroki:
  - Do każdej wersji językowej artykułu należy wyekstrahować miary jakości (patrz rozdziały nr 5 „Miary oraz wymiary jakości artykułów Wikipedii” oraz nr 7 „Miary oraz wymiary jakości infoboksu”). Szczególną uwagę należy zwrócić uwagę na miary, który wykazały wysoką ważność w modelach jakości (patrz rozdziały „Budowanie modeli jakości artykułów” oraz „Budowanie modeli jakości infoboksów”).
  - Ocena jakości wersji językowych wybranego artykułu oraz infoboksów (J1-J5). Oceny te będą występować jako dodatkowe miary.

W celu przedstawienia działania tej metody na konkretnych przykładach rozpatrzymy artykuł o Poznaniu w wersji ARPUB. Zróbmy to na postawie wyżej opisanych kroków:

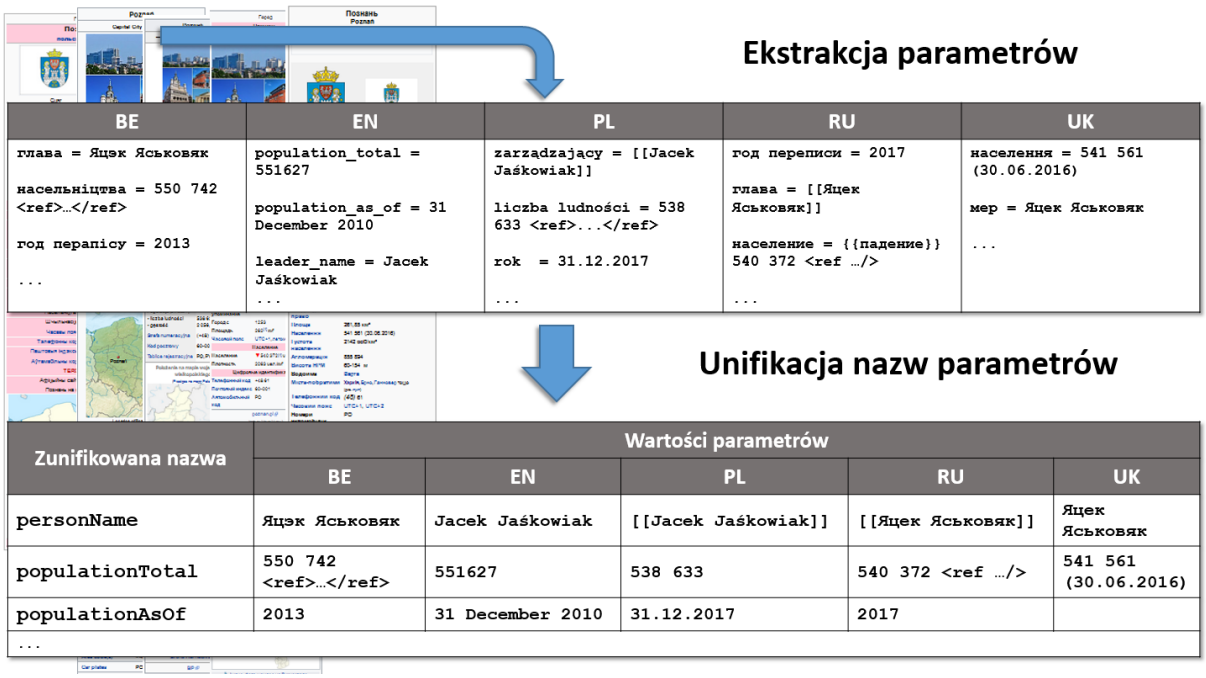
- Zostało wybranych 5 wersji językowych (angielska, rosyjska, polska, ukraińska, białoruska) dla artykułu o Poznaniu[2]. We wszystkich rozpatrywanych językach artykuł posiada infoboks podobnego typu. Np. w wersji polskiej ma on nazwę „Polskie miasto infobox”.

[2] https://pl.wikipedia.org/wiki/Poznań

- Dla każdej wersji językowej tego artykułu zostały wyekstrachowane miary jakości artykułu (A1-A138) oraz miary jakości infoboksu (I1-I45).
- Na podstawie powyższych miar wyznaczono miary jakości z modeli jakości artykułów (J1-J5).
- Zostały wyekstrahowane parametry infoboksów w każdej wersji językowej. Ogólny schemat ekstrakcji pokazany na rys. 9.4.
- Używając mapowań DBpedii zostały zunifikowane nazwy parametrów. Ogólny schemat unifikacji nazw pokazany na rys. 9.4.
- Używając technik dostępnuch w ramach DBpedia Framework[3] oraz własnych algorytmów, zostały zunifikowane wartości parametrów.
  - W celu unifikacji parametru „personName" w wersji rosyjskiej zostały użyte semantyczne powiązania. W wersji ukraińskiej wartość była przekształcona na podstawie podobieństwa z wersją rosyjską, która posiada semantyczne powiązanie z innymi nazwami. W przypadku białoruskiej wersji zostało wykorzystano podobieństwo z wersją rosyjską oraz zasady białoruskiej ortografii oraz interpunkcji (Ustawa, 2008).
  - Wartości parametru „populationTotal" zostały zunifikowane w taki sposób, aby występowała tylko wartość liczbowa, którą można łatwo porównać.
  - Wartości parametru „populationAsOf", które wskazują aktualność danych o populacji („populationTotal"), zostały przekształcone w ujednolicony format daty. Jeżeli był wpisany tylko rok, to został dopisany ostatni dzień oraz miesiąc. W przypadku wersji ukraińskiej, wartość tego parametry była przeniesiona z innego parametru („populationTotal").

  Ogólny schemat unifikacji wartości parametrów infoboksów został pokazany na rys. 9.5.
- Porównanie wartości parametrów w wybranych językach.
  - „personName": we wszystkich wersjach językowych jest zgodność co do wartości.
  - „populationTotal": każda wersja językowa posiada inną wartość. Innymi słowy - nie ma zgodności pomiędzy wersjami językowymi.
  - „populationAsOf": prawie każda wersja językowa posiada różną wartość tego parametru, oprócz wartości które wspólnie posiadają polska oraz rosyjska wersja.
- Ocena wersji językowych na podstawie modeli jakości infoboksów w wersji rozszerzonej. Największą liczbę punktów zdobyła wersja polska. Oznacza to, że w przypadku braku

[3] https://github.com/dbpedia/extraction-framework



BE	EN	PL	RU	UK
глава = Яцэк Яськовяк насельніцтва = 550 742 <ref>…</ref> год перапісу = 2013 ...	population_total = 551627 population_as_of = 31 December 2010 leader_name = Jacek Jaśkowiak ...	zarządzający = [[Jacek Jaśkowiak]] liczba ludności = 538 633 <ref>...</ref> rok = 31.12.2017 ...	год переписи = 2017 глава = [[Яцек Яськовяк]] население = {{падение}} 540 372 <ref …/> ...	населення = 541 561 (30.06.2016) мер = Яцек Яськовяк ...

Zunifikowana nazwa	Wartości parametrów				
	BE	EN	PL	RU	UK
personName	Яцэк Яськовяк	Jacek Jaśkowiak	[[Jacek Jaśkowiak]]	[[Яцек Яськовяк]]	Яцек Яськовяк
populationTotal	550 742 <ref>…</ref>	551627	538 633	540 372 <ref …/>	541 561 (30.06.2016)
populationAsOf	2013	31 December 2010	31.12.2017	2017	
...					

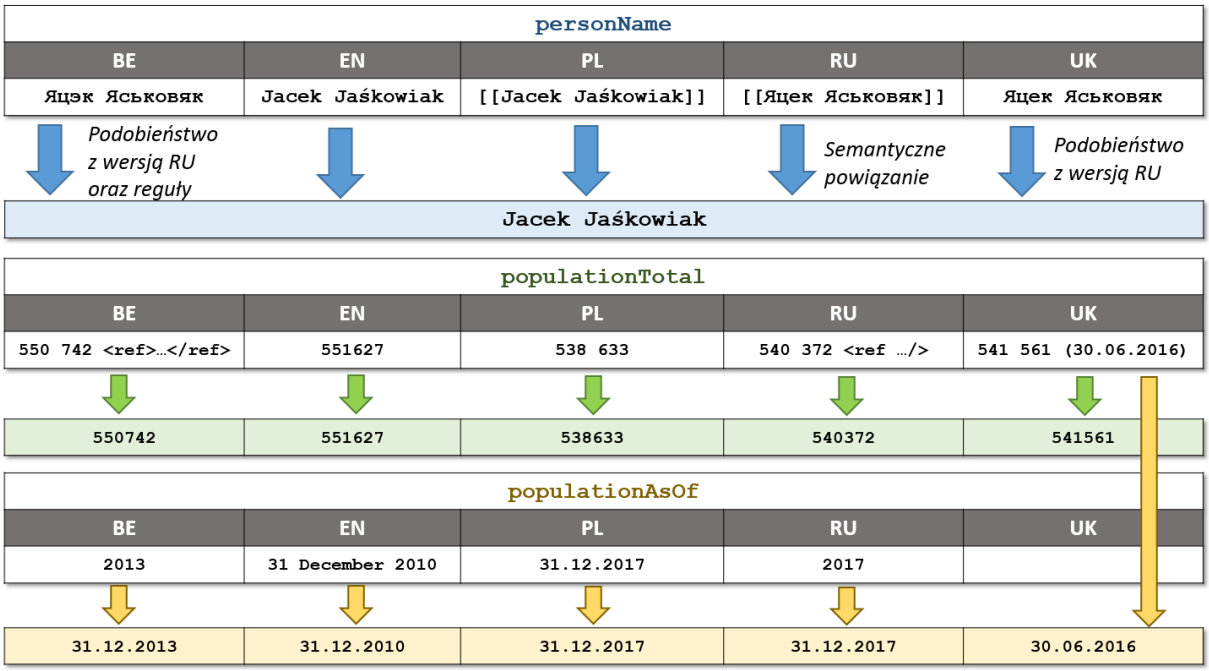
Rysunek 9.4. Schemat ekstrakcji infoboksów oraz unifikacji nazw parametrów na przykładzie artykułu o Poznaniu w wersji ARPUB.

Źródło: Opracowanie własne.

pewnego parametru infoboksu w innych wersjach językowych, do jego uzupełnienia preferowana jest wartość w wersji polskiej. Jeżeli niektóre wersje mają spójne wartości parametru, obliczana jest liczba punktów za jakość uzyskana przez model. W związku z tym, dla parametru „populationTotal" wartość z wersji PL będzie najlepsza. W przypadku parametru „populationAsOf", mamy do czynienia ze zgodnością pomiędzy wersjami PL oraz RU. Wiemy, że wersja PL posiada w naszym przykładzie maksymalną liczbę punktów za jakość, w związku z tym dodatkowa zgodność „wzmacnia" wybór wartości z tych wersji.

9.4 Podsumowanie

W niniejszym rozdziale została przedstawiona metoda porównywania wielojęzycznych informacji w Wikipedii. Ważnym etapem tej metody jest unifikacja wartości parametrów na podstawie określonych reguł oraz semantycznych powiązań obiektów. W celu unifikacji parametrów infoboksów były wykorzystane mapowania dostępne w ramach serwisu DBpedia Mappings. Proces mapowania parametrów można również zautomatyzować oraz usprawnić za pomocą różnych metod, opisanych w niniejszym rozdziale.



Rysunek 9.5. Schemat unifikacji wartości wybranych parametrów na przykładzie artykułu o Poznaniu.

Źródło: Opracowanie własne.

Drugim ważnym elementem metody jest analiza jakości źródła, w którym został umieszczony infoboks. W tym przypadku był wykorzystany model do obliczenia punktów za jakość infoboksu przygotowany w ramach rozprawy. W przypadku braku spójności na poziomie konkretnego parametru pomiędzy wersjami językowymi, najlepsza wartość była wyznaczana na podstawie liczby punktów lub sumy liczby punktów, jeżeli ta wartość była użyta w dwóch lub większej liczbie wersji językowych.

Opisana w niniejszym rozdziale metoda porównywania wielojęzycznych informacji na podstawie analizy jakości może pomóc we wzbogaceniu mniej rozwiniętych wersji językowych Wikipedii. W następnym rozdziale zostanie zaproponowana taka metoda wzbogacenia.

Rozdział 10

Metoda wzbogacenia informacji

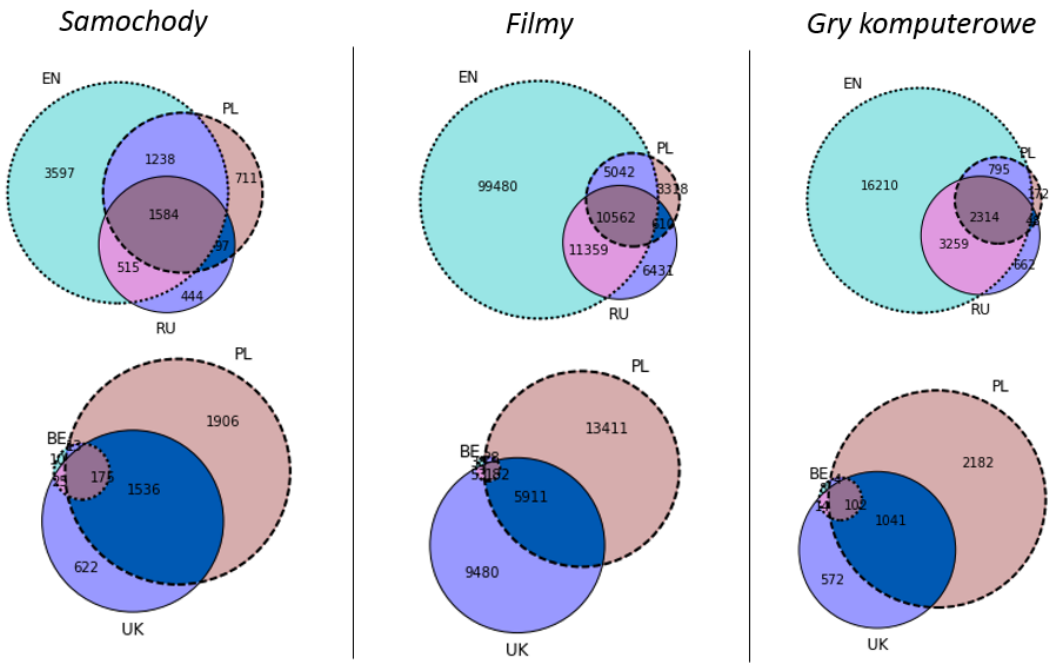
W tym rozdziale opisano metodę wzbogacenia informacji na podstawie metody porównywania wielojęzycznych informacji oraz modelu jakości infoboksów, które zostały opisane wcześniej.

10.1 Wprowadzenie

Pomimo istnienia możliwości dodawania artykułów w różnych językach na ten sam temat (jako odpowiednik artykułów w innych językach), czasem można zauważyć istotne różnice pomiędzy liczbą opisanych obiektów oraz ich reprezentacji w każdej wersji językowej. Pokrycie artykułów na wybrane tematy w różnych wersjach językowych zostało przedstawione na rysunku 10.1). Wynika z niego, że białoruska wersja, jako najmniejsza, posiada jednocześnie największy potencjał, jeśli chodzi o możliwość wzbogacenia informacji na podstawie innych wersji językowych Wikipedii.

Automatyczne wzbogacenie informacji w takich popularnych źródłach jak Wikipedia nie jest zupełnie nowym zagadnieniem. Istnieją różne sposoby na przenoszenia danych pomiędzy wersjami językowymi (Kaffee, 2016). Wśród nich są mapowania DBpedii, które były opisane w poprzednim rozdziale. Do wzbogacania artykułów Wikipedii mogą być również wykorzystywane Wikidane (Sáez i Hogan, 2018).

Istnieją również algorytmy (boty), które już przenoszą treści pomiędzy wersjami językowymi, dokonując na przykład automatycznego tłumaczenia. W związku z tym zdarza się, że liczba artykułów w danej wersji językowej Wikipedii nie ma związku z liczbą redaktorów lub osób posługujących się tym językiem. Na przykład Wikipedia w języku cebuańskim z ponad 5 milionami artykułów formalnie zajmuje drugie miejsce (po angielskiej) w rankingu największych



Rysunek 10.1. Pokrycie tematów w różnych wersjach językowych Wikipedii.

Źródło: Obliczenia własne.

wersji językowych Wikipedii (Wikipedia Meta-Wiki, 2018b). Język cebuański jest używany na Filipinach, a ogólna liczba osób posługujących się tym językiem wynosi do 24% populacji tego kraju (Oard i in., 2003). Inny przykład - szwedzka Wikipedia, która posiada ponad 3.7 mln artykułów (Wikipedia Meta-Wiki, 2018b), co pozwala jej zajmować trzecie miejsce w rankingu największych wersji językowych Wikipedii. Na świecie jest około 10.5 mln osób mówiących w języku szwedzkim (The Network to Promote Linguistic Diversity (NPLD), 2018). Autorem większości artykułów w tych dwóch wersjach językowych jest bot, który ma nazwę Lsjbot[1].

Należy zauważyć, że praktycznie żadne z istniejących rozwiązań nie bierze pod uwagę porównanie jakości pomiędzy wersjami językowymi, z których potencjalnie może zostać wyekstrahowana informacja do późniejszego przenoszenia na inne wersję językowe. Opisana w tym rozdziale metoda jest propozycją sekwencji czynności, które należy przeprowadzić przed rozpoczęciem przenoszenia danych do różnych wersji językowych Wikipedii.

[1] https://en.wikipedia.org/wiki/Lsjbot

10.2 Metoda wzbogacenia informacji

Celem metody jest przenoszenie informacji pomiędzy wersjami językowymi biorąc pod uwagę jakość tych informacji oraz źródła, z którego one pochodzą. Głównym obiektem przenoszenia są dane z infoboksów. Źródłem tych infoboksów są artykuły, które posiadają różne miary jakości.

Podstawowym zadaniem metody jest przenoszenie danych infoboksu o najlepszej jakości do wersji innych, mniej rozwiniętych wersji językowych Wikipedii. Mniej rozwiniętą w tym kontekście będziemy nazywać wersję językową, która nie posiada artykułu na określony temat lub posiada artykuł bez infoboksu.

W celu wzbogacenia informacji zaproponowano następujące kroki:

- Pierwszym etapem w ramach przedstawionej metody jest identyfikacja artykułów lub tematów wraz z wersjami językowymi, w ramach których będzie przeprowadzona analiza jakości oraz późniejsze przenoszenie danych.
- Po identyfikacji artykułów oraz wersji językowych Wikipedii, należy przeprowadzić analizę liczebności wersji językowych dla każdego z tematów oraz każdej wersji językowej. Celem danego etapu jest zidentyfikowanie wersji językowych źródłowych oraz wersji docelowych. Źródłowe wersje to takie, które posiadają infoboks z wypełnionym co najmniej jednym parametrem. Takie infoboksy to potencjalne źródła dla wzbogacenia innych (docelowych) wersji językowych.
- Następnym etapem jest porównanie jakości źródłowych wersji językowych przy pomocy przygotowanych w ramach danej rozprawy modeli. Wynikiem tego etapu musi być zidentyfikowanie najlepszej źródłowej wersji językowej, która będzie wykorzystana do przenoszenia danych do docelowej (docelowych) wersji językowych.
- Dla wersji docelowych trzeba zidentyfikować nazwę infoboksu, która może być odpowiednikiem infoboksu z najlepszej źródłowej wersji językowej. Do znalezienia takiego odpowiednika można używać semantyczne powiązania pomiędzy stronami szablonów przy użyciu interwikilinks.
- Biorąc pod uwagę reguły unifikacji parametrów infoboksów, należy zidentyfikować wypełnione parametry w infoboksie najlepszej wersji źródłowej oraz ich odpowiedniki w infoboksie wersji docelowej. Unifikacja parametrów infoboksów może być przeprowadzona na podstawie mapowań DBpedii oraz innych technik.

- Przed przenoszeniem parametrów, należy przeprowadzić analizę ich wartości. Na przykład, jeżeli wartość parametru zawiera link do obiektu, który jest opisany w Wikipedii jako osobny artykuł, trzeba sprawdzić czy istnieje nazwa tego obiektu w języku docelowym. Jeżeli tak - należy zmienić wartość na odpowiednią. Jeżeli nie - można stosować różne techniki, w zależności od specyfik oraz różnic pomiędzy wersją źródłową oraz docelową. Musi zostać przeprowadzona również analiza szablonów, które używane są w wartościach parametrów infoboksu. Przede wszystkim należy zbadać, czy odpowiednie szablony są opisane w docelowej wersji językowej oraz sprawdzić spójność nazewnictwa parametrów.
- Po przygotowaniu kodu infoboksu z wypełnionymi wartościami parametrów zgodnie z zasadami docelowej wersji językowej, należy wstawić dany kod w odpowiednie miejsce artykułu (zazwyczaj to jest początek artykułu).
- Jeżeli język docelowy nie posiada artykułu, należy utworzyć artykuł z nazwą, która jest odpowiednikiem (tłumaczeniem) nazwy z wersji źródłowej.

10.3 Zbiór danych

Do celów niniejszego rozdziału zostały wybrane artykuły z 4 wersji językowych Wikipedii: angielska (EN), polska (PL), rosyjska (RU) oraz ukraińska (UK). Białoruska wersja (BE) została wybrana do wzbogacenia informacji.

W celu przeprowadzenia dalszych eksperymentów, podobnie jak było to zrobione w poprzednim rozdziale, zostały dobrane do próby artykuły z infoboksami w wersji ARPU opisujące miasta Polski, miasta Ukrainy, miasta Rosji, firmy, uniwersytety oraz gry komputerowe. Z tego zbioru zostały usunięte artykuły w wersji ARPUB, czyli te, które posiadają dodatkowo wersję białoruską. Szczegółowe informacje odnośnie zbiorów wersji ARPU oraz ARPUB zostały przedstawione we Wprowadzeniu do rozdziału 8 „Budowanie modeli jakości infoboksów", liczebności tych zbiorów podane w tabeli 8.1.

W wyniku, po usunięciu 1654 artykułów, zostało 4441 art., które posiadają cztery wersji językowe (angielska, rosyjska, polska, ukraińska) oraz które będą przedmiotem dalszej analizy. Dla każdej wersji językowej tych artykułów zostały wyekstrahowane miary jakości. Po tym, wartości tych miar zostały znormalizowane podobnie jak to było opisane w rozdziale „Budowanie modeli jakości infoboksów".

Zatem wynikiem eksperymentów na tym zbiorze danych będzie utworzenie 4441 artykułów z infoboksami w wersji docelowej (białoruskiej) na różne tematy.

Dla wszystkich artykułów zostały wyekstrahowane miary jakości artykułów oraz infoboksów, opisane w niniejszej pracy. Dane do analizy były ekstrahowane na podstawie kopii zapasowych Wikipedii w lipcu 2018 roku oraz serwisów internetowych, które były wcześniej opisane w rozprawie.

10.4 Lokalna wersja Wikipedii

Eksperymenty po wzbogaceniu danych były przeprowadzone na lokalnej wersji Wikipedii, która została zaimplementowana na podstawie ogólnodostępnych kopii zapasowych Wikipedii przy użyciu ogólnodostępnego oprogramowania MediaWiki.

Wolne oprogramowanie MediaWiki pierwotnie było opracowano dla Wikipedii, ale obecnie jest wykorzystywane przez wiele innych serwisów wiki (Kaffee, 2016). Do uruchomienia tego oprogramowania został wykorzystany XAMPP, który jest wieloplatformowym, zintegrowanym pakietem, pozwalającym na uruchomienie własnego serwera WWW na komputerze do obsługi dynamicznych stron.

Oprócz MediaWiki, do uruchomienia lokalnej wersji Wikipedii należy zainstalować dodatkowe rozszerzenia, m.in to Scribunto (pozwala na osadzanie języków skryptowych na stronach wiki), Lua (do uruchomienia kodu w języku Lua) oraz inne.

Następnie zostały zaimportowane pliki kopii zapasowej Wikipedii wybranej wersji językowej do bazy danych. W naszym przypadku - to białoruska Wikipedia[2].

10.5 Eksperymenty

W celu przeprowadzenia eksperymentów zostały określone nowe artykuły, które można utworzyć na podstawie artykułów z każdego zbioru danych, gdzie każdy artykuł posiada wszystkie 4 wersje językowe. Na przykład, jeżeli rozpatrzymy zbiór danych z polskimi miastami, to dla wzbogacenia białoruskiej Wikipedii może być utworzono 701 nowych artykułów z infoboksem na podstawie analizy jakości w artykułach z 4 wersji językowych. Liczba utworzonych artykułów w białoruskiej wersji Wikipedii w ramach każdego zbioru danych jest podana w tabeli 10.1.

[2]Kopie zapasowe dla białoruskiej Wikipedii dostępne na stronie `https://dumps.wikimedia.org/bewiki/`

Tabela 10.1. Liczba utworzonych artykułów w białoruskiej wersji Wikipedii w ramach każdego zbioru danych.

Zbór	Liczba nowych artykułów
Miasta Polski	701
Miasta Rosji	87
Miasta Ukrainy	1001
Firmy	1262
Uniwersytety	392
Gry komputerowe	998
Razem	**4441**

Źródło: Opracowanie własne.

W przypadku artykułów dotyczących miast Polski, po określeniu liczby potencjalnych nowych artykułów (701) w ramach zbioru danych, należało zidentyfikować nazwę infoboksu, który będzie podobny do infoboksów z innych wersji językowych oraz który będzie związany z tym tematem. Dla białoruskiej wersji istnieje bezpośredni odpowiednik szablonu, który można znaleźć przy pomocy interwikilinks na stronie opisu infoboksu „Polskie miasto infobox" w języku polskim.

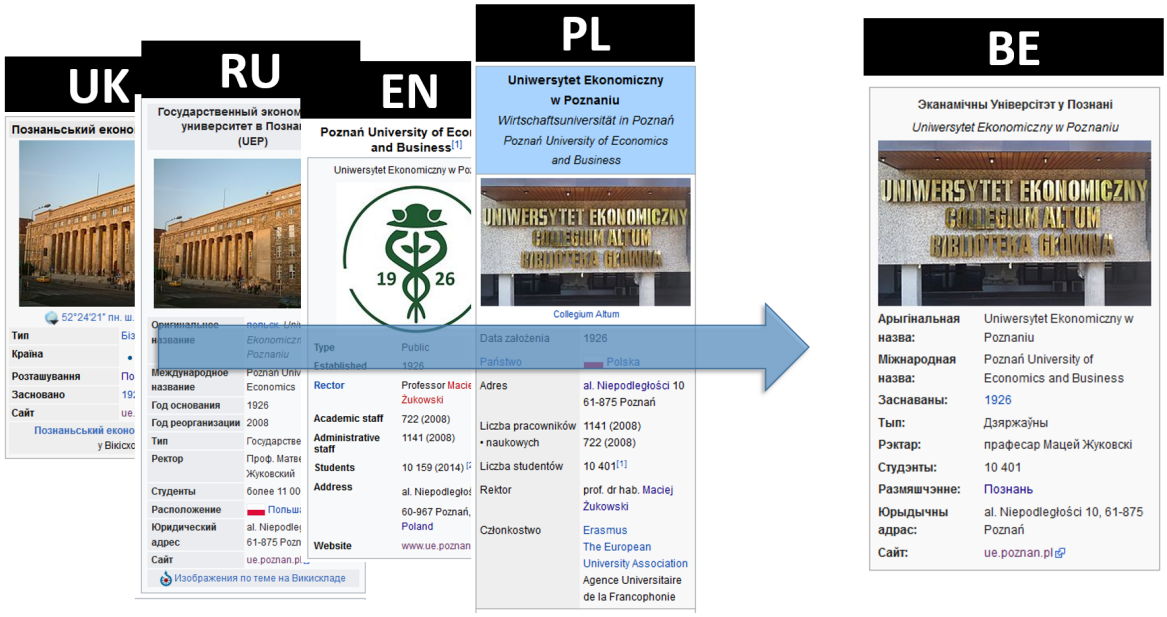
Utworzonych artykułów o miastach Rosji jest stosunkowo mało w ogólnym zbiorze danych. Jest to związane m.in. z tym, że angielska Wikipedia stosuje oddzielny szablon dla miast Rosji pod nazwą „Infobox Russian inhabited locality". Oddzielny szablon stosuje również wersja rosyjska oraz ukraińska. W przypadku wersji białoruskiej oraz polskiej, stosują one ogólny infoboks dla miast z Rosji (np. w polskiej - „Miejscowość infobox").

Podobnie jak i w poprzednich zbiorach danych, po określeniu liczby potencjalnych nowych artykułów (1001) w ramach zbioru danych o miastach Ukrainy, został zidentyfikowany infoboks w ukraińskiej Wikipedii. W tym przypadku angielska, polska oraz białoruska wersja Wikipedii, w odróżnieniu od ukraińskiej oraz rosyjskiej, stosują ogólny infoboks, który opisuje miasta z różnych państw.

Infoboksy o firmach mają swoje odpowiedniki we wszystkich rozpatrywanych wersjach językowych. Nowych artykułów w ramach tej tematyki utworzono 1262.

W przypadku uniwersytetów istnieją oddzielne infoboksy na ten temat w każdym z rozpatrywanych języków. Ogólna liczba nowych artykułów o uniwersytetach wynosi 392. Przykład wzbogacenia białoruskiej Wikipedii o infoboks opisujący Uniwersytet Ekonomiczny w Poznaniu został pokazany na rys. 10.2.

W białoruskiej Wikipedii stosunkowo słabo rozwinięty jest temat gier komputerowych, w związku z tym udział nowych artykułów w tym przypadku jest największy. Ogólnie zostało utworzone 998 artykułów z infoboksami na ten temat.



Rysunek 10.2. Przykład wzbogacenia białoruskiej Wikipedii o infoboks opisujący Uniwersytet Ekonomiczny w Poznaniu.

Źródło: Opracowanie własne.

Jak już było wspomniano wcześniej, tłumaczenie, przekształcenie oraz przenoszenie informacji pomiędzy różnymi wersjami językowymi Wikipedii było już tematem innych prac (**Upadhyay2018**; Bizer i in., 2009; Lehmann i in., 2015; T.-N. Nguyen i in., 2016). Natomiast nowatorską oraz najważniejszą częścią składową tej metody jest automatyczny dobór wersji językowych z informacjami o najwyższej jakości z uwzględnieniem ponad 100 różnych miar. W związku z tym, w następnym rozdziale została opisana weryfikacja tej metody po kątem prawidłowości doboru najlepszych wersji językowych, co pozwala na późniejsze przenoszenia informacji wysokiej jakości.

10.6 Podsumowanie

W niniejszym rozdziale została przedstawiona metoda wzbogacenia informacji na podstawie analizy jakości. Zostały przeprowadzone eksperymenty na konkretnych przykładach. Ta metoda to propozycja działań, które należy wykonać w celu przeniesienia danych do różnych wersji językowych Wikipedii.

Samą decyzję o automatycznym wzbogaceniu musi podjąć uprawniony użytkownik czy społeczność użytkowników konkretnej wersji językowej Wikipedii.

Ważniejszą częścią składową metody wzbogacenia informacji w ramach tej rozprawy jest porównywanie informacji wielojęzycznych na podstawie analizy ich jakości. W następnym rozdziale została przeprowadzona weryfikacja tej metody.

Rozdział 11

Ewaluacja metod

Celem rozdziału jest omówienie wyników eksperymentów przeprowadzonych dla oceny skuteczności działania opracowanych metod. Ewaluacja ma udowodnić tezę przyjęta dla pracy i sformułowaną we Wstępie, a także pozwolić na osiągniecie niektórych celów szczegółowych rozprawy.

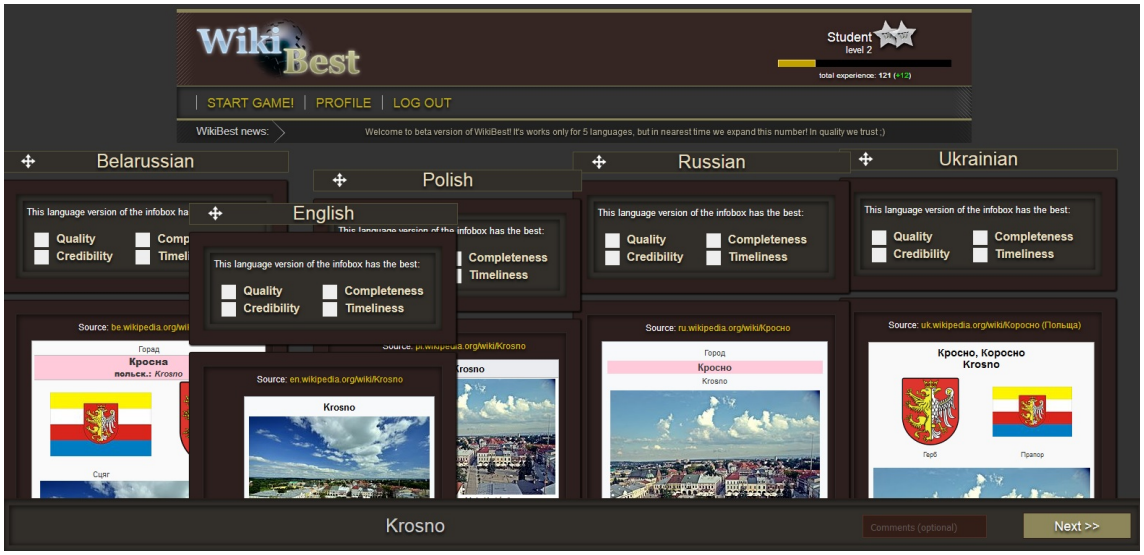
11.1 Wprowadzenie

W celu weryfikacji przedstawionej metody zostały zaproszone osoby, które posiadają znajomość co najmniej czterech języków: polski, angielski, rosyjski, ukraiński. Dodatkowo były zaproszone osoby, które rozumieją język białoruski. Dlatego eksperci byli podzieleni na dwie grupy, w zależności od znajomości języków:

- Grupa podstawowa: osoby, które zadeklarowały znajomość języka polskiego, angielskiego, rosyjskiego oraz ukraińskiego;
- Grupa zaawansowana: osoby, które zadeklarowały znajomość języka polskiego, angielskiego, rosyjskiego, ukraińskiego oraz białoruskiego.

11.2 Narzędzie do zbierania danych od ekspertów

W celu usprawnienia zbierania danych od ekspertów został utworzony serwis WikiBest (adres WWW: `https://www.wikibest.net`). Za pośrednictwem tego serwisu eksperci mogli wybierać najlepsze wersje językowy infoboksów na określone tematy. W zależności od grupy, na



Rysunek 11.1. Interfejs serwisu WikiBest - na przykładzie pokazane infoboksy na temat miasta Krosno w 5 wersjach językowych.

Źródło: `https://wikibest.net`

ekranie po kolei pojawiały się infoboksy opisujące ten sam obiekt (np. miasto) w różnych wersjach językowych (patrz rys. 11.1).

Każde zadanie polegało na tym, że spośród przedstawionych na ekranie wersji językowych użytkownicy musieli zaznaczyć, która wersja językowa w ich opinii posiada dane najwyższej jakości. Dodatkowo, eksperci mogli zaznaczyć najlepsze infoboksy pod kątem kompletności, wiarygodności oraz aktualności. Zadania były dobrane w taki sposób, aby co najmniej 5 ekspertów miało możliwość oceny tego samego tematu.

Eksperci byli zapraszani za pośrednictwem informacji umieczonych na różnych publicznych serwisach internetowych, takich jak Facebook, Reddit, Twitter, Vk.com, Habr.com, DobreProgramy.pl, Medium.com. Informacja była opublikowana m.in. na następujących stronach (dane na wrzesień 2018 r.):

- Społeczność „Wikipedia” na portalu Reddit `https://www.reddit.com/r/wikipedia/`. Ta strona posiada ponad 220 tys. subskrydentów.
- Artykuł na rosyjskim portalu ”Habr„ - `https://habr.com/post/418713/` . Ta wiadomość posiada ponad 4 tyś. wyświetleń.
- Oficjalna strona „WikiData” w serwisie społecznościowym Twitter - `https://twitter.com/wikidata` . Ta strona posiada ponad 13 tys. obserwujących.
- Strona WikiRank w serwisie Twitter - `https://twitter.com/infoboxesnet`. Strona posiada ponad 600 obserwujących.

- Grupa „pl.wikipedia” https://www.facebook.com/groups/165267100158576/ - posiada ponad 300 uczestników.
- Grupa „Wikipedia Weekly” https://www.facebook.com/groups/wikipediaweekly/ - posiada ponad 1300 uczestników.
- Inne, w tym tematyczne strony projektów o Wikipedii w różnych serwisach społecznościowych.

Przed rozpoczęciem pracy w serwisie WikiBest należało się zarejestrować. Przy rejestracji były pobierane m.in. następujące dane: Rok urodzenia, Państwo, Wykształcenie, Płeć, Znajomość języków. Każdy z ekspertów zostawiał również dane kontaktowe.

W okresie badawczym w serwisie zostały zarejestrowane 74 osoby. Z tego grona zostały wybrane osoby, które uczestniczyły w rozwiązaniu co najmniej 1 zadania oraz osoby, których średni czas rozwiązania każdego zadania nie był mniejszy niż 10 sekund. Dodatkowo zostały wyeliminowane osoby, które nieprawidłowo oceniły co najmniej połowę tematów z listy kontrolnej. Zadania z listy kontrolnej pojawiały się przypadkowo, bez ujawniania użytkownikowi szczególnej ważności takich zadań.

W wyniku powyższej weryfikacji do ewaluacji metody zostały wybrane 52 osoby, z których 7 wykazały znajomość wszystkich pięciu języków (grupa zaawansowana), a pozostałe 45 osób znajomością czterech języków (grupa podstawowa).

11.3 Opinie ekspertów

W związku z tym, że każdy temat był oceniony przez 5 ekspertów, w ramach zbioru danych ARPU liczącego 6095 artykułów, zostały uwzględnione 30475 głosów ekspertów w ramach wszystkich rozpatrywanych tematów. Na przykład na zbiorze danych dotyczących polskich miast w 4 wersjach językowych otrzymano 4505 głosów na 901 artykułów. Liczba głosów na poszczególne wersje językowe w ramach artykułów o polskich miastach rozkładała się następująco: polska - 4345, angielska - 125, rosyjska - 26, ukraińska - 4. Tabela 11.1 przedstawia liczebności głosów ekspertów na poszczególne w wersje językowe artykułów w ramach wszystkich rozpatrywanych tematów.

Tabela 11.1. Liczba głosów ekspertów na poszczególne w wersje językowe artykułów w ramach rozpatrywanych tematów.

Temat	Liczba głosów na wersje językowe			
	EN	PL	RU	UK
Miasta Polski	38	4453	12	2
Miasta Rosji	397	0	4024	4
Miasta Ukrainy	101	73	2470	3876
Firmy	5908	170	761	96
Uniwersytety	2298	197	373	97
Gry komputerowe	3902	164	1024	35
Razem	**12644**	**5057**	**8664**	**4110**

Źródło: Opracowanie własne na podstawie danych od ekspertów.

11.4 Zgodność ekspertów

Do pomiaru zgodności pomiędzy ekspertami będzie użyty współczynnik Kappa Fleissa (Fleiss, 1971) oraz Kappa Randolpha (Randolph, 2005). W odróżnieniu od innych pomiarów zgodności (np. współczynnika Kappa Cohena), przy obliczaniu współczynnika Kappy Fleissa każdy z n losowo wybranych obiektów może być oceniany przez inny losowy zestaw k sędziów.

Jak już było wspomniano wcześniej, zadania do oceny były dobrane w taki sposób, aby co najmniej 5 ekspertów miało możliwość oceny tego samego artykułu. Analiza współczynnika zgodności będzie opierać się zatem na danych przekształconych do tabeli o n wierszach i c kolumnach, gdzie c stanowi liczbę możliwych kategorii, do których eksperci (czy sędziowie) przydzielają najlepsze wersje językowe poszczególnych artykułów. W wyniku tego każde przecięcie x_{ij} takiej tabeli będzie pokazywało liczbę sędziów wydających określone w kolumnie j opinie na temat i artykułu. Przykład takiej tabeli jest podany w 11.2. W tym przykładzie dla artykułu nr 636 mamy następujące opinie: dwaj eksperci uważają, że infoboks najlepszej jakości posiada polska wersja, 2 innych ekspertów odznaczyli ukraińską wersję jako najlepszą, a jeden ekspert uważa że infoboks tego artykułu jest najlepszej jakości w angielskiej Wikipedii. Natomiast co do artykułu nr 15 wszyscy eksperci mają wspólną opinię - wersją językową z infoboksem o najlepszej jakości jest polska Wikipedia.

Współczynnik Kappa Fleissa ($\hat{\kappa}_F$) wyrażony będzie wzorem:

$$\hat{\kappa}_F = \frac{P_o - P_e}{1 - P_e} \tag{11.1}$$

Tabela 11.2. Przykład tabeli z liczbą głosów za poszczególną wersję językową dla określonych artykułów Wikipedii do obliczenia współczynnika zgodności Kappa Fleissa.

Nr artykułu	Wersja językowa			
	EN	PL	RU	UK
7	2	3	0	0
15	0	5	0	0
449	2	2	1	0
558	0	3	2	0
636	1	2	0	2

Źródło: Opracowanie własne na podstawie danych od ekspertów.

gdzie:

$P_o = \frac{1}{kn(k-1)} \sum_{i=1}^{n} \sum_{j=1}^{c} x_{ij} - kn$,

$P_e = \sum_{i=1}^{c} q_j^2$,

$q_j = \frac{1}{km} \sum_{i=1}^{n} x_{ij}$.

Jeżeli wartość $\hat{\kappa}$ jest równa 1, to oznacza pełną zgodność sędziów. W przypadku $\hat{\kappa}$ równej 0, stwierdza się opinie ekspertów wydane były w sposób losowy. Wartości ujemne danego współczynnika wskazują na zgodność mniejszą niż na poziomie losowym.

Dla zbioru danych dotyczących miast polskich w 4 wersjach językowych $\hat{\kappa}_F = 0.27$. Wartość $\hat{\kappa}$ w przedziale od 0.2 do 0.4 oznacza dostateczne (fair) porozumienie pomiędzy ekspertami (Landis i Koch, 1977; Viera, Garrett i in., 2005).

Współczynnik Kappa Fleissa był przedstawiony w 1971 roku i był nową metodą obliczenia zgodności więcej niż dwóch ekspertów. Jednak takie podejście pasuje bardziej dla badań zgodności, gdzie mamy do czynienia ze stałym (czy ustalonym) rozkładem brzegowym opinii, kiedy eksperci znają a priori liczbę przypadków, które należy przypisać do poszczególnej kategorii (Randolph, 2005). W związku z tym, w innych warunkach przy wysokiej zgodności ekspertów możemy mieć niską wartość współczynnika Kappa Fleissa . To dotyczy m.in. sytuacji, kiedy eksperci nie są zmuszani do przypisania z góry ustalonej liczby przypadków do określonej kategorii (w naszym przypadku - wersji językowej) (Brennan i Prediger, 1981). Dlatego, dodatkowo będzie obliczony współczynnik Kappa Randolpha, który został przedstawiony w 2005 roku jako alternatywa współczynniku Kappa Fleissa (Randolph, 2005; Warrens, 2010).

Współczynnik Kappa Randolpha obliczany według podobnego ogólnego wzoru, jednak ze zmianą zasady obliczenia parametru P_e:

$$\hat{\kappa}_R = \frac{P_o - P_e}{1 - P_e} \tag{11.2}$$

gdzie:

$P_o = \frac{1}{kn(k-1)} \sum_{i=1}^{n} \sum_{j=1}^{c} x_{ij} - kn$,

$P_e = 1/c$.

Dla tego samego zbioru danych współczynnik Kappa Randolpha jest równy około 0.95, co wskazuje na prawie idealne porozumienie pomiędzy ekspertami (Landis i Koch, 1977; Viera, Garrett i in., 2005).

Podobnie jak i do tematu z miastami Polski zostały obliczone współczynniki dla innych tematów. W tabeli 11.3 pokazano wartości współczynników Kappa Fleissa oraz Kappa Randolpha dla rozpatrywanych tematów.

Tabela 11.3. Wartości współczynników Kappa Fleissa oraz Kappa Randolpha dla rozpatrywanych tematów.

Temat	Kappa Fleissa	Kappa Randolpha
Miasta Polskie	0,33	0,98
Miasta Ukrainy	0,52	0,76
Miasta Rosji	0,33	0,89
Firmy	0,64	0,91
Uniwersytety	0,78	0,92
Gry komputerowy	0,55	0,83

Źródło: Obliczenia własne na podstawie danych od ekspertów.

11.5 Ewaluacja metody

Do ewaluacji metody będzie wykorzystany współczynnik Kappa Cohena, który może ocenić stopień zgodności (rzetelności) pomiędzy dwoma sędziami. W tym przypadku w roli sędziów będą występowali eksperci (jako pierwszy) oraz model jakości (jako drugi), zbudowany w ramach rozprawy. Z próby w 6095 artykułów zostało wyeliminowano 68 przypadków (co stanowi ok. 1% próby), gdzie eksperci nie mieli wystarczającego poziomu zgodności. W związku z tym, do analizy wybrano 6027 artykułów, gdzie dla każdego przypadku co najmniej 3 z 5 ekspertów wybrali tą samą wersję językową jako najlepszą.

Tabela 11.4 przedstawia liczbę najlepszych wersji językowych nadanych przez model w wersji podstawowej oraz przez ekspertów do obliczenia współczynnika Kappa Cohena

Tabela 11.4. Liczba najlepszych wersji językowych nadanych przez model w wersji podstawowej oraz przez ekspertów do obliczenia współczynnika Kappa Cohena.

		Model podstawowy				
		EN	**PL**	**RU**	**UK**	**Suma**
Eksperci	**EN**	**2425**	36	87	13	**2561**
	PL	5	**972**	2	3	**982**
	RU	133	15	**1298**	216	**1662**
	UK	29	3	49	**741**	**822**
	Suma	**2592**	**1026**	**1436**	**973**	**6027**

Źródło: Obliczenia własne oraz na podstawie danych od ekspertów.

Wartość współczynnika Kappa Cohena w tym przypadku wynosi 0,8603. Ta wartość wskazuje na bardzo dobrą zgodność pomiędzy ocenami ekspertów oraz modelem jakości, zaproponowanym w rozprawie.

Tabela 11.5 przedstawia liczbę najlepszych wersji językowych nadanych przez model w wersji rozszerzonej oraz przez ekspertów do obliczenia współczynnika Kappa Cohena

Tabela 11.5. Liczba najlepszych wersji językowych nadanych przez model w wersji rozszerzonej oraz przez ekspertów do obliczenia współczynnika Kappa Cohena.

		Model rozszerzony				
		EN	**PL**	**RU**	**UK**	**Suma**
Eksperci	**EN**	**2500**	3	57	1	**2561**
	PL	12	**968**	0	2	**982**
	RU	93	1	**1424**	144	**1662**
	UK	3	1	25	**793**	**822**
	Suma	**2608**	**973**	**1506**	**940**	**6027**

Źródło: Obliczenia własne oraz na podstawie danych od ekspertów.

Wartość współczynnika Kappa Cohena w tym przypadku wynosi 0,9189. Zatem można wnioskować, że model rozszerzony jest bardziej precyzyjny niż model podstawowy również na innych danych uczących, które nie były brane pod uwagę podczas budowania modelu.

11.6 Podsumowanie

W niniejszym rozdziale został przedstawiony sposób weryfikacji metody zaproponowanej w rozprawie. W szczególności zostało opisane narzędzie WikiBest, które umożliwia zbieranie informacji przez Internet od ekspertów na temat porównania jakości danych w infoboksach Wikipedii w różnych językach.

Ewaluacja pokazała wysoki stopień zgodności ekspertów w procesie oceny jakości infoboksów w rożnych wersjach językowych. Największą zgodność ocen eksperckich z modelem jakości infoboksów można uzyskać przy stosowaniu wersji rozszerzonej modelu, które oprócz miar infoboksów zawiera również miary artykułów Wikipedii. W tym przypadku osiągnięto wartość 0,9189 współczynnika Kappa Cohena, co pokazuje wysoką korelację pomiędzy ocenami ekspertów a opracowanym w ramach rozprawy modelem oceny jakości infoboksów.

Rozdział 12

Podsumowanie

Przedmiotem prezentowanej rozprawy było wzbogacenie wielojęzycznych informacji w serwisach wiki na podstawie analizy ich jakości. Zagadnienie to może pełnić istotną rolę w funkcjonowaniu podmiotów gospodarczych oraz osób prywatnych, ponieważ ilość i jakość informacji w dużym stopniu decydują o jakości podejmowanych decyzji w różnych gałęziach gospodarki.

Prezentowana praca miała na celu pokazania nowego podejścia do rozwiązania znanych problemów.

12.1 Wkład pracy

Głównym celem pracy było opracowanie metody porównywania oraz wzbogacania informacji w wielojęzycznych serwisach wiki na podstawie analizy ich jakości na przykładzie Wikipedii.

Główny cel pracy miał zostać osiągnięty poprzez realizacje pięciu celów szczegółowych. Te cele to:

- Opracowanie metody automatycznej oceny jakości strony wiki w różnych językach z wykorzystaniem odpowiednich miar.
- Opracowanie metody porównania jakości infoboksu z jakością strony wiki.
- Opracowanie metody identyfikacji infoboksów oraz parametrów w nim umieszczonych o najwyższej jakości spośród odpowiedników strony wiki w różnych wersjach językowych.
- Opracowanie metody wzbogacenia infoboksów pomiędzy wersjami językowymi wiki z wykorzystaniem semantycznej reprezentacji elementów tych infoboksów.
- Opracowanie metody tworzenia nowej strony w określonej wersji językowej z wybranymi elementami infoboksu o najwyższej jakości z innych wersji językowych wiki.

Każdy z tych celów został osiągnięty poprzez wypracowanie odpowiednich artefaktów oraz przeprowadzenie ich ewaluacji:

- Nowe miary jakości artykułów oraz infoboksów serwisów wiki. W szczególności, miara syntetyczna - do oceny jakości artykułów w skali od 0 do 100. Miara pokazała wyższą ważność w otrzymanej metodzie niż tradycyjne oceny jakości, stosowane w ramach Wikipedii.
- Model oceny jakości infoboksów w wersji podstawowej oraz wersji rozszerzonej. Ten model wykazał szczególną ważność miar dotyczących popytu na informację.
- Metoda porównywania wielojęzycznych informacji w Wikipedii na podstawie semantycznych powiązań oraz oceny jakości.

Osiągając przedstawione cele, wykazano tezę pracy: **Metoda oceny jakości strony wykorzystująca semantyczne powiązania z innymi wersjami językowymi oraz uwzględniająca popyt na informację pozwala na porównanie i wzbogacenie informacji w serwisach wiki.**

W trakcie prac służących osiągnięciu celu głównego pracy i wykazaniu jej tezy osiągnięto również poboczne rezultaty, które również stanowią wkład do analizowanej dziedziny:

- Opracowanie nowych miar do oceny jakości artykułów Wikipedii oraz infoboksów. W tym, opracowanie miary syntetycznej do oceny jakości artykułów w różnych wersjach językowych Wikipedii.
- Typologizacja miar jakości oraz wymiarów jakości artykułów Wikipedii.
- Opracowanie modeli estymacji jakości artykułów w różnych językach na podstawie modeli określonej wersji językowej z użyciem dychotomicznej lub nominalnej zmiennej zależnej.
- Analiza współzależności miar infoboksów oraz artykułów.
- Opracowanie modeli oceny jakości infoboksów w różnych wersjach językowych.
- Opracowanie metody porównywania wielojęzycznych informacji na podstawie analizy ich jakości.
- Opracowanie metody wzbogacenia wielojęzycznych informacji na podstawie analizy ich jakości.

12.2 Dalsze badania

Dalsze badania mogą się toczyć wokół zdefiniowania i ekstrakcji nowych miar jakości. Takie miary mogą się pojawić w wyniku wprowadzenia nowych możliwości systemów zarządzania serwisami wiki oraz rozwoju serwisów zewnętrznych, które mogą oceniać różne elementy artykułów serwisów wiki (np. referencje).

Kolejnym kierunkiem dalszych badań może być zmiana podejścia do doboru zmiennej zależnej. Tutaj mogą być różne możliwości, które będą uniwersalne dla każdego lub wybranych wersji językowych. Szczególnie obiecującym jest rozwój wskaźnika syntetycznego, do którego można wprowadzić dodatkowe miary wraz z przypisaniem wag do każdej z nich, w zależności od ważności w konkretnej wersji językowej.

W ramach rozprawy zostały pokazane eksperymenty na jednym z popularnych przykładów serwisów wiki - Wikipedii. W związku z tym wyniki prezentowanej pracy mogą być wykorzystane również do wzbogacenia innych wielojęzycznych serwisów wiki, działających na podobnych zasadach.

Miary związane z referencjami pokazały wysoką ważność w modelach jakości artykułów oraz infoboksów. To wynika m.in. z faktu, że użytkownicy Wikipedii szczególnie dbają o weryfikowalność przedstawionych informacji w artykułach. W związku z tym przyszłe badania mogą się skupiać wokół analizy jakości źródeł Wikipedii. Można tutaj włączyć różne narzędzia, np. bazy bibliograficzne: Microsoft Academic Graph, AMiner[1], Altmetric[2] oraz inne. Narzędzia mogą pomoc również w określeniu rodzaju źródła (np. czy to jest praca naukowa, jak często cytowana w innych pracach etc.).

Szczególnie interesującym mogą być również badanie neutralności treści artykułów Wikipedii. W tym celu mogą być użyte metody analizy wydźwięku (Małyszko, 2013).

Oddzielne badania mogą być poświęcone miarom lingwistycznym i ich wpływowi na jakość artykułów. Takie miary mogą pozwolić na ekstrakcje faktów, co umożliwi porównanie poprawności ze źródłem lub z innymi wersjami językowymi.

[1] https://www.openacademic.ai/oag/
[2] https://www.altmetric.com/

Bibliografia

Abramowicz, W. (2008). *Filtrowanie informacji*. Poznań: Wydawnictwo Akademii Ekonomicznej w Poznaniu.

Adler, B. T. & De Alfaro, L. (2007). A content-driven reputation system for the wikipedia. *Proceedings of the 16th international conference on World Wide Web WWW 07*, *7*(Generic), 261. doi:10.1145/1242572.1242608

Aganbegjan, A. G. (2012). Dostizhenie vysshego urovnja prodolzhitelnosti zhizni v Rossii. *Rossijskoe predprinimatelstvo*, (2).

Ahmeti, A., Fernández, J. D., Polleres, A. & Savenkov, V. (2017). Updating Wikipedia via DBpedia Mappings and SPARQL. W E. Blomqvist, D. Maynard, A. Gangemi, R. Hoekstra, P. Hitzler i O. Hartig (Red.), *The Semantic Web* (s. 485–501). Cham: Springer International Publishing.

Alexa. (2018). wikia.com Traffic Statistics.

Aljumaili, M., Karim, R. & Tretten, P. (2016). Quality of Streaming Data in Condition Monitoring Using ISO 8000. W U. Kumar, A. Ahmadi, A. K. Verma i P. Varde (Red.), *Current Trends in Reliability, Availability, Maintainability and Safety* (s. 703–715). Cham: Springer International Publishing.

Anderka, M. (2013). *Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia* (PhD, Bauhaus-Universitaet Weimar Germany).

Arazy, O. (2010). Determinants of Wikipedia Quality : the Roles of Global and Local Contribution Inequality. *New York*, 233–236. doi:10.1145/1718918.1718963

Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M. & Gergle, D. (2012). Omnipedia: bridging the wikipedia language gap. W *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (s. 1075–1084). ACM.

Bartosik-Purgat, M., Mruk, H. & Schroeder, J. (2012). *Dostępność i wykorzystanie informacji o rynkach i partnerach zagranicznych w procesie internacjonalizacji polskich przedsiębiorstw*. Wydawnictwo Uniwersytetu Ekonomicznego.

Batini, C. & Scannapieco, M. (2016). Data Quality Dimensions. W *Data and Information Quality: Dimensions, Principles and Techniques* (s. 21–51). doi:10.1007/978-3-319-24106-7_2

Bednarek-Michalska, B. (2007). Ocena jakości informacji elektronicznej. Pułapki sieci.

Benson, P. (2008). NATO codification system as the foundation for ISO 8000, the international standard for data quality. *Oil IT Journal*.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. & Hellmann, S. (2009). DBpedia-A crystallization point for the Web of Data. *Web Semantics: science, services and agents on the world wide web*, *7*(3), 154–165.

Blumenstock, J. E. (2008a). *Automatically Assessing the Quality of Wikipedia Articles*. doi:10.1080/17439880802324251

Blumenstock, J. E. (2008b). Size matters: word count as a measure of quality on wikipedia. W *WWW* (s. 1095–1096). doi:10.1145/1367497.1367673

Boehm, B. W. (1988). A spiral model of software development and enhancement. *Computer*, *21*(5), 61–72.

Bormuth, J. R. (1966). Readability: A new approach. *Reading research quarterly*, 79–132.

Boruszewski, J. (2012). Jakość i wiarygodność informacji w infobrokerstwie. *Lingua ac Communita*, *22*, 241–250.

Bould, M. D., Hladkowicz, E. S., Pigford, A.-A. E., Ufholz, L.-A., Postonogova, T., Shin, E. & Boet, S. (2014). References that anyone can edit: review of Wikipedia citations in peer reviewed health science literature. *BMJ*, *348*, g1585.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, California: Wadsworth International Group.

Brennan, R. L. & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and psychological measurement*, *41*(3), 687–699.

Buecheler, T., Sieg, J. H., Füchslin, R. M. & Pfeifer, R. (2010). Crowdsourcing, Open Innovation and Collective Intelligence in the Scientific Method-A Research Agenda and Operational Framework. W *ALIFE* (s. 679–686).

Callahan, E. S. & Herring, S. C. (2011). Cultural bias in Wikipedia content on famous persons. *Journal of the Association for Information Science and Technology*, *62*(10), 1899–1915.

Caylor, J. S. & Sticht, T. G. (1973). Development of a Simple Readability Index for Job Reading Material.

Chen, H.-H. (2012). How to Use Readability Formulas to Access and Select English Reading Materials. *Journal of Educational Media & Library Sciences*, *50*(2).

Chistiakov, S. (2013). Sluchajnye lesa: obzor. *Trudy Karelskogo nauchnogo centra RAN*, *1*, 117–136.

Coleman, M. & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*(2), 283.

Commission of the European Communities. (2002). eEurope 2002: Quality criteria for health related websites. doi:10.2196/jmir.4.3.e15

Conti, R., Marzini, E., Spognardi, A., Matteucci, I., Mori, P. & Petrocchi, M. (2014). Maturity assessment of Wikipedia medical articles. W *Computer-Based Medical Systems (CBMS), 2014 IEEE 27th International Symposium on* (s. 281–286). IEEE.

Crawford, H. (2001). Encyclopedias. *Reference and information services: An introduction*, 433–459.

Dale, E. & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational research bulletin*, 37–54.

Dalip, D. H., Gonçalves, M. A., Cristo, M. & Calado, P. (2017). A general multiview framework for assessing the quality of collaboratively created content on web 2.0. *Journal of the Association for Information Science and Technology*, *68*(2), 286–308.

Dalip, D. H., Gonçalves, M. A., Cristo, M. & Calado, P. (2009). Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. W *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries* (s. 295–304). doi:10.1145/1555400.1555449

Dalip, D. H., Gonçalves, M. A., Cristo, M. & Calado, P. (2011). Automatic Assessment of Document Quality in Web Collaborative Digital Libraries. *Journal of Data and Information Quality*, *2*(3), 1–30. doi:10.1145/2063504.2063507

Dang, Q.-V. & Ignat, C.-L. (2016a). Measuring Quality of Collaboratively Edited Documents: The Case of Wikipedia. W *Collaboration and Internet Computing (CIC), 2016 IEEE 2nd International Conference on* (s. 266–275). IEEE.

Dang, Q.-V. & Ignat, C.-L. (2016b). Quality assessment of Wikipedia articles without feature engineering. W *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)* (s. 27–30).

de Freitas, C. (2003). Tourism climatology: evaluating environmental information for decision making and business planning in the recreation and tourism sector. *international Journal of Biometeorology*, *48*(1), 45–54.

Deng, Z. & Luo, L. (2007). An exploratory discuss of new ways for competitive intelligence on Web2. 0. W *Integration and Innovation Orient to E-Society Volume 2* (s. 597–604). Springer.

di Sciascio, C., Strohmaier, D., Errecalde, M. & Veas, E. (2017). WikiLyzer: interactive information quality assessment in Wikipedia. W *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (s. 377–388). ACM.

Drèze, X. & Zufryden, F. (2004). Measurement of online visibility and its impact on Internet traffic. *Journal of interactive marketing*, *18*(1), 20–37.

English Wikipedia. API sandbox. (nodate).

Eppler, M. (2013). *Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes*. Springer Berlin Heidelberg.

Eppler, M. J. (2006). *Managing information quality: Increasing the value of information in knowledge-intensive products and processes*. Springer Science & Business Media.

Experian QAS. (2013). *The Data Advantage: How accuracy creates opportunity* (tech. rep. Nr February). Experian QAS.

Fandom. (2018). Explore.

Färber, M., Bartscherer, F., Menne, C. & Rettinger, A. (2016). Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. *Semantic Web*, (Preprint), 1–53.

Ferschke, O., Gurevych, I. & Rittberger, M. (2012). FlawFinder: A Modular System for Predicting Quality Flaws in Wikipedia. W *CLEF (Online Working Notes/Labs/Workshop)* (s. 1–10).

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

Flekova, L., Ferschke, O. & Gurevych, I. (2014). What makes a good biography?: multidimensional quality analysis based on wikipedia article feedback data. W *Proceedings of the 23rd international conference on World wide web* (s. 855–866). ACM.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, *32*(3), 221.

Freund, Y. & Mason, L. (1999). The alternating decision tree learning algorithm. W *Proceeding of the Sixteenth International Conference on Machine Learning* (Vol. 99, s. 124–133). doi:10.1093/jxb/ern164

Freund, Y. & Schapire, R. E. (1996). Experiments with a new boosting algorithm. W *Thirteenth International Conference on Machine Learning* (s. 148–156). San Francisco: Morgan Kaufmann.

Friedman, J., Hastie, T. & Tibshirani, R. (2000). Additive logistic regression : A statistical view of boosting. *Annals of statistics*, *28*(2), 337–407.

Gama, J. (2004). Functional Trees. *55*(3), 219–250.

Garvin, D. (1984). What Does ”Product QualityŻeally Mean? *26*, 25–43.

Ge, M. & Helfert, M. (2008). Data and information quality assessment in information manufacturing systems. W *11th International Conference, BIS 2008, Innsbruck, Austria, May 5-7, 2008.* (s. 380–389). doi:10.1007/978-3-540-79396-0_33

Giles, J. (2005). Internet encyclopaedias go head to head. Nature Publishing Group.

Goodman, A. (2008). Winning Results with Google AdWords.

Grantner, E. (2007). ISO 8000: a standard for data quality. *Logistics Spectrum*, *41*(4).

Greenfield, G. R. (1999). *Classic Readability Formulas in an EFL Context: Are They Valid for Japanese Speakers?* (Doctoral dissertation, Temple University).

Grijzenhout, S. & Marx, M. (2013). The quality of the XML Web. *Journal of Web Semantics*, *19*, 59–68. doi:10.1016/j.websem.2012.12.001

Grudzień, Ł. (2012). Koncepcja oceny jakości informacji o procesach w systemach zarządzania. *Konferencja IZIP Zakopane*.

Gunning, R. (1952). The technique of clear writing.

Halfaker, A. (2017). Interpolating quality dynamics in wikipedia and demonstrating the keilana effect. W *Proceedings of the 13th International Symposium on Open Collaboration* (s. 19). ACM.

Halfaker, A., Kraut, R. & Riedl, J. (2009). A Jury of Your Peers : Quality, Experience and Ownership in Wikipedia. *WikiSym'09*, 1–10. doi:10.1145/1641309.1641332

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10–18.

Härting, R.-C., Mohl, M., Steinhauser, P. & Möhring, M. (2016). Search Engine Visibility Indices Versus Visitor Traffic on Websites. W *International Conference on Business Information Systems* (s. 91–101). Springer.

Hastie, T. & Tibshirani, R. (1998). Classification by Pairwise Coupling. W M. I. Jordan, M. J. Kearns i S. A. Solla (Red.), *Advances in Neural Information Processing Systems* (Vol. 10), MIT Press.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). Unsupervised learning. W *The elements of statistical learning* (s. 485–585). Springer.

Heinrich, B. & Klier, M. (2015). Metric-based data quality assessment — Developing and evaluating a probability-based currency metric. *Decision Support Systems*, *72*, 82–96. doi:10.1016/j.dss.2015.02.009

Herrera-Viedma, E., Pasi, G., Lopez-Herrera, A. G. & Porcel, C. (2006). Evaluating the information quality of web sites: A methodology based on fuzzy computing with words. *Journal of the American Society for Information Science and Technology*, *57*(4), 538–549.

Hevner, A. R. (2004). DESIGN SCIENCE IN INFORMATION SYSTEMS RESEARCH. *MIS Quarterly*, *28*(1), 75–105.

Holmes, G., Donkin, A. & Witten, I. H. (1994). Weka: A machine learning workbench. W *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on* (s. 357–361). IEEE.

Holmes, G., Pfahringer, B., Kirkby, R., Frank, E. & Hall, M. (2001). Multiclass alternating decision trees. W *ECML* (s. 161–172). Springer.

Horn, C., Zhila, A., Gelbukh, A., Kern, R. & Lex, E. (2013). Using factual density to measure informativeness of Web documents. W *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22-24; 2013; Oslo University; Norway. NEALT Proceedings Series 16* (085, s. 227–238). Linköping University Electronic Press.

Hu, M., Lim, E.-P., Sun, A., Lauw, H. W. & Vuong, B.-Q. (2007). Measuring article quality in wikipedia. W *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management - CIKM '07* (s. 243–252). doi:10.1145/1321440.1321476

Hube, C., Fischer, F., Jäschke, R., Lauer, G. & Thomsen, M. R. (2017). World Literature According to Wikipedia: Introduction to a DBpedia-Based Framework. *arXiv preprint arXiv:1701.00991*.

International Telecommunication Union. (2017). Measuring the Information Society Report 2017, Volume 1.

Internet World Stats. (2018). World Internet Users Statistics and 2018 World Population Stats.

ISO/TS. (2011). *Technical specification ISO/TS 8000-150:2011(E)—data quality—Part 1: overview*. Geneva, Switzerland.

Jackson, S. L. (2014). *Research methods: A modular approach*. Cengage Learning.

Jang, S., Megawati, J. C., Choi, J. & Yi, M. Y. (2015). Semi-Automatic Quality Assessment of Linked Data without Requiring Ontology. W *NLP-DBPEDIA@ ISWC* (s. 45–55).

Jemielniak, D. (2013). Życie wirtualnych dzikich. *Netnografia Wikipedii, najwiekszego projektu współtworzonego przez ludzi. Warszawa: Poltext*.

Jennex, M. E. & Bartczak, S. E. (2013). A revised knowledge pyramid. *International Journal of Knowledge Management (IJKM)*, *9*(3), 19–30.

Jiao, Y.-y. & Yuan, J. (2008). Research on the Collective Knowledge Sharing and Innovation Service Based on WIKI [J]. *Information Science*, *5*, 003.

Kaffee, L. A. (2016). *Generating article placeholders from Wikidata for Wikipedia: increasing access to free and open knowledge* (Doctoral dissertation, Hochschule für Technik und Wirtschaft Berlin).

Kane, G. C. (2011). A multimethod study of information quality in wiki collaboration. *ACM Transactions on Management Information Systems (TMIS)*, *2*(1), 4.

Keerthi, S., Shevade, S., Bhattacharyya, C. & Murthy, K. (2001). Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*, *13*(3), 637–649.

Kidholm, K., Ølholm, A. M., Birk-Olsen, M., Cicchetti, A., Fure, B., Halmesmäki, E., Kahveci, R., Kiivet, R.-A., Wasserfallen, J.-B., Wild, C. i in. (2015). Hospital managers' need for information in decision-making–An interview study in nine European countries. *Health Policy*, *119*(11), 1424–1432.

Killoran, J. B. (2013). How to use search engine optimization techniques to increase website visibility. *IEEE Transactions on professional communication*, *56*(1), 50–66.

Kim, E. & Choi, K.-S. (2014). Cross-lingual property alignment for DBpedia ontology using triple conceptualization. W *Proceedings of the 13th International Semantic Web Conference: Poster*.

Kim, S. & Stoel, L. (2004). Apparel retailers: website quality dimensions and satisfaction. *Journal of Retailing and Consumer Services*, *11*(2), 109–117.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L. & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Naval Technical Training Command Millington TN Research Branch.

Kittur, A. & Kraut, R. E. (2008). Harnessing the wisdom of crowds in wikipedia. *Proceedings of the ACM 2008 conference on Computer supported cooperative work - CSCW '08*, 37. doi:10.1145/1460563.1460572

Klusch, M. (2001). Information agent technology for the internet: A survey. *Data & Knowledge Engineering*, *36*(3), 337–372.

Kohavi, R. (1995). The Power of Decision Tables. W *8th European Conference on Machine Learning* (s. 174–189). Springer.

Kohavi, R. (1996). Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. W *Second International Conference on Knoledge Discovery and Data Mining* (s. 202–207).

Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R. & Zaveri, A. (2014). Test-driven evaluation of linked data quality. W *Proceedings of the 23rd international conference on World Wide Web* (s. 747–758). ACM.

Kousha, K. & Thelwall, M. (2017). Are wikipedia citations important evidence of the impact of scholarly articles and books? *Journal of the Association for Information Science and Technology*, *68*(3), 762–779.

Kumar, L. & Kumar, N. (2014). SEO Technique for a website and its effectiveness in context of Google Search Engine. *International Jouranl of Computer Science and Engineering (IJCSE) Vol*, *2*, 113–118.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Landwehr, N., Hall, M. & Frank, E. (2005). Logistic Model Trees. *Machine Learning*, *95*(1-2), 161–205.

Lee, Y. W. (2003). Crafting rules: context-reflective data quality problem solving. *Journal of Management Information Systems*, *20*(3), 93–119.

Lehmann, J., Müller-Birn, C., Laniado, D., Lalmas, M. & Kaltenbrunner, A. (2014). Reader preferences and behavior on wikipedia. W *Proceedings of the 25th ACM conference on Hypertext and social media* (s. 88–97). ACM.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S. i in. (2015). DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, *6*(2), 167–195.

Lerner, J. & Lomi, A. (2018). Knowledge categorization affects popularity and quality of Wikipedia articles. *PloS one*, *13*(1), e0190674.

Lewoniewski, W. (2017a). Completeness and Reliability of Wikipedia Infoboxes in Various Languages. W W. Abramowicz (Red.), *Business Information Systems Workshops* (s. 295–305). Cham: Springer International Publishing.

Lewoniewski, W. (2017b). Enrichment of Information in Multilingual Wikipedia Based on Quality Analysis. W *International Conference on Business Information Systems* (s. 216–227). Springer.

Lewoniewski, W., Härting, R.-C., Węcel, K., Reichstein, C. & Abramowicz, W. (2018). Application of SEO Metrics to Determine the Quality of Wikipedia Articles and Their Sources. W *The 24th International Conference on Information and Software Technologies (ICIST 2018)*. (in press).

Lewoniewski, W., Kasprzak, A., Węcel, K. & Abramowicz, W. (2018). Kompletność danych o produktach w infoboksach różnych wersji językowych Wikipedii. W *VII Ogólnopolska Konferencja Naukowa. Matematyka i informatyka na usługach ekonomii im. Profesora Zbigniewa Czerwińskiego*. (w publikacji).

Lewoniewski, W., Khairova, N., Węcel, K., Stratiienko, N. & Abramowicz, W. (2017). Using Morphological and Semantic Features for the Quality Assessment of Russian Wikipedia. W R. Damaševičius i V. Mikayte (Red.), *Information and Software Technologies: 23rd International Conference, ICIST 2017, Druskininkai, Lithuania, October 12–14, 2017, Proceedings* (s. 550–560). doi:10.1007/978-3-319-67642-5_46

Lewoniewski, W., Wecel, K. & Abramowicz, W. (2017). Determining Quality of Articles in Polish Wikipedia Based on Linguistic Features. DOI.

Lewoniewski, W. & Węcel, K. (2017). Relative Quality Assessment of Wikipedia Articles in Different Languages Using Synthetic Measure. W W. Abramowicz (Red.), *Business Information Systems Workshops: BIS 2017 International Workshops, Poznań, Poland, June 28-30, 2017, Revised Papers* (s. 282–292). doi:10.1007/978-3-319-69023-0_24

Lewoniewski, W., Węcel, K. & Abramowicz, W. (2015). Analiza porównawcza modeli jakości informacji w narodowych wersjach Wikipedii. W T. Porębska-Miąc (Red.), *Systemy Wspo-*

magania Organizacji SWO 2015 (s. 133–154). Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

Lewoniewski, W., Węcel, K. & Abramowicz, W. (2016). Quality and Importance of Wikipedia Articles in Different Languages. W *Information and Software Technologies: 22nd International Conference, ICIST 2016, Druskininkai, Lithuania, October 13-15, 2016, Proceedings* (s. 613–624). doi:10.1007/978-3-319-46254-7_50

Lewoniewski, W., Węcel, K. & Abramowicz, W. (2017a). Analysis of References Across Wikipedia Languages. W R. Damasevicius i V. Mikayte (Red.), *Information and Software Technologies: 23rd International Conference, ICIST 2017, Druskininkai, Lithuania, October 12–14, 2017, Proceedings* (s. 561–573). doi:10.1007/978-3-319-67642-5_47

Lewoniewski, W., Węcel, K. & Abramowicz, W. (2017b). Relative Quality and Popularity Evaluation of Multilingual Wikipedia Articles. *Informatics*, *4*(4). doi:10.3390/informatics4040043

Lewoniewski, W., Węcel, K. & Abramowicz, W. (2017c). Zagadnienia jakości w crowdsourcingu. *Outsourcing Magazine*, *2017*(4 (42)), 52–55.

Lex, E., Voelske, M., Errecalde, M., Ferretti, E., Cagnina, L., Horn, C., Stein, B. & Granitzer, M. (2012). Measuring the quality of web content using factual information. *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality - WebQuality '12*, 7. doi:10.1145/2184305.2184308

Lih, A. (2004). Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource. *5th International Symposium on Online Journalism*, 31.

Lin, J. & Fenner, M. (2014). An analysis of Wikipedia references across PLOS publications. W *altmetrics14: Expanding impacts and metrics An ACM Web Science Conference 2014 Workshop* (s. 23–26).

Lipka, N. & Stein, B. (2010). Identifying Featured Articles in Wikipedia: Writing Style Matters. *Proceedings of the 19th International Conference on World Wide Web (2010)*, 1147–1148. doi:10.1145/1772690.1772847

Liu, J. & Ram, S. (2018). Using big data and network analysis to understand Wikipedia article quality. *Data & Knowledge Engineering*.

Lucassen, T. & Schraagen, J. M. (2010). Trust in Wikipedia: how users trust information from an unknown source. W *Proceedings of the 4th workshop on Information credibility* (s. 19–26). ACM.

Luyt, B. & Tan, D. (2010). Improving Wikipedia's credibility: References and citations in a sample of history articles. *Journal of the Association for Information Science and Technology*, *61*(4), 715–722.

Łapczyński, M. (2003). Drzewa klasyfikacyjne w badaniach satysfakcji i lojalności klientów. *StatSoft Polska*, 93–102.

Madnick, S., Wang, R., Lee, Y. & Zhu, H. (2009). Overview and Framework for Data and Information Quality Research. *ACM Journal of Data and Information Quality*, *1*(1), 1–22. doi:10.1145/1515693.1516680.http

Madnick, S. & Zhu, H. (2006). Improving data quality through effective use of data semantics. *Data & Knowledge Engineering*, *59*(2), 460–475. doi:10.1016/j.datak.2005.10.001

Małyszko, J. (2013). Aspect-Aware Identification of Opinion Phrases Polarity Based on Summaries of Consumer Opinions about Products and Services. W W. Abramowicz (Red.), *Business Information Systems Workshops* (s. 278–289). Berlin, Heidelberg: Springer Berlin Heidelberg.

Maynes, R. & Everdell, I. (2014). The evolution of Google search results pages and their effects on user behaviour.

Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of reading*, *12*(8), 639–646.

Mihindukulasooriya, N., Rico, M., García-Castro, R. & Gómez-Pérez, A. (2015). An analysis of the quality issues of the properties available in the Spanish DBpedia. W *Conference of the Spanish Association for Artificial Intelligence* (s. 198–209). Springer.

Mohanty, R., Seth, D. & Mukadam, S. (2007). Quality dimensions of e-commerce and their implications. *Total Quality Management & Business Excellence*, *18*(3), 219–247.

Moyer, D., Carson, S. L., Dye, T. K., Carson, R. T. & Goldbaum, D. (2015). Determining the influence of Reddit posts on Wikipedia pageviews. W *Ninth International AAAI Conference on Web and Social Media* (s. 75–82). AAAI Press Oxford, UK.

Natale, D. (2011). Complexity and data quality. W *Poster e Atti Conferenza* (s. 13–16).

Nguyen, N., Cao, D. & Nguyen, A. (2018). Automatically Mapping Wikipedia Infobox Attributes to DBpedia Properties for Fast Deployment of Vietnamese DBpedia Chapter. W N. T.

Nguyen, D. H. Hoang, T.-P. Hong, H. Pham i B. Trawiński (Red.), *Intelligent Information and Database Systems* (s. 127–136). Cham: Springer International Publishing.

Nguyen, T., Moreira, V., Nguyen, H., Nguyen, H. & Freire, J. (2011). Multilingual Schema Matching for Wikipedia Infoboxes. *Proc. VLDB Endow. 5*(2), 133–144. doi:10 . 14778 / 2078324.2078329

Nguyen, T.-N., Takeda, H., Nguyen, K., Ichise, R. & Cao, T.-D. (2016). Type Prediction for Entities in DBpedia by Aggregating Multilingual Resources. W *International Semantic Web Conference (Posters & Demos)*.

Nielsen, F. Å. (2007). Scientific citations in Wikipedia. *arXiv preprint arXiv:0705.2106*.

Oard, D. W., Doermann, D., Dorr, B., He, D., Resnik, P., Weinberg, A., Byrne, W., Khudanpur, S., Yarowsky, D., Leuski, A. i in. (2003). Desparately seeking cebuano. W *Companion volume of the proceedings of HLT-NAACL 2003-short papers*.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web.* Stanford InfoLab.

Page, R. D. (2010). Wikipedia as an encyclopaedia of life. *Organisms Diversity & Evolution*, *10*(4), 343–349.

Palmero Aprosio, A., Giuliano, C. & Lavelli, A. (2013). Towards an Automatic Creation of Localized Versions of DBpedia. W H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. Noy, C. Welty i K. Janowicz (Red.), *The Semantic Web – ISWC 2013* (s. 494–509). Berlin, Heidelberg: Springer Berlin Heidelberg.

Parssian, A., Sarkar, S. & Jacob, V. S. (2004). Assessing data quality for information products: impact of selection, projection, and Cartesian product. *Management Science*, *50*(7), 967–982.

Paulheim, H. (2017). Data-Driven Joint Debugging of the DBpedia Mappings and Ontology. W E. Blomqvist, D. Maynard, A. Gangemi, R. Hoekstra, P. Hitzler i O. Hartig (Red.), *The Semantic Web* (s. 404–418). Cham: Springer International Publishing.

Platt, J. (1998). Fast Training of Support Vector Machines using Sequential Minimal Optimization. W B. Schoelkopf, C. Burges i A. Smola (Red.), *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

Podshivalenko, G. (2010). Investicionnyj klimat i investicionnaja privlekatelnost. *Finansovaja analitika: problemy i reshenija*, (15).

Polska Wikipedia. (2018). Pomoc:Przestrzeń nazw.

Price, R. & Shanks, G. (2016). A semiotic information quality framework: development and comparative analysis. W *Enacting Research Methods in Information Systems* (s. 219–250). Springer.

Pun, J. C. & Lochovsky, F. H. (2004). Ranking Search Results by Web Quality Dimensions. *J. Web Eng. 3*(3-4), 216–235.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers.

Randolph, J. J. (2005). Free-Marginal Multirater Kappa (multirater K [free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa. *Online submission*.

Ransbotham, S. & Kane, G. C. (2011). Membership turnover and collaboration success in online communities: Explaining rises and falls from grace in Wikipedia. *Mis Quarterly*, 613–627.

Ransbotham, S., Kane, G. C. & Lurie, N. H. (2012). Network characteristics and the value of collaborative user-generated content. *Marketing Science*, *31*(3), 387–405.

Reinoso, A. J. (2011). *Temporal and behavioral patterns in the use of Wikipedia* (Doctoral dissertation, Universidad Rey Juan Carlos). http://gsyc.es/ ajreinoso/thesis/.

Rico, M., Mihindukulasooriya, N., Kontokostas, D., Paulheim, H., Hellmann, S. & Gómez-Pérez, A. (2018). Predicting Incorrect Mappings: A Data-driven Approach Applied to DBpedia. W *Proceedings of the 33rd Annual ACM Symposium on Applied Computing* (s. 323–330). SAC '18. doi:10.1145/3167132.3167164

Rinser, D., Lange, D. & Naumann, F. (2013). Cross-lingual entity matching and infobox alignment in Wikipedia. *Information Systems*, *38*(6), 887–907.

RYTE GmbH. (2018). Search Engine Optimization.

Sáez, T. & Hogan, A. (2018). Automatically Generating Wikipedia Info-boxes from Wikidata. W *Companion Proceedings of the The Web Conference 2018* (s. 1823–1830). WWW '18. doi:10.1145/3184558.3191647

Saito, T. & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, *10*(3), e0118432.

Schaal, M., Smyth, B., Mueller, R. M. & MacLean, R. (2012). Information quality dimensions for the social web. W *Proceedings of the International Conference on Management of Emergent Digital EcoSystems* (s. 53–58). ACM.

Schroeder, B. (2007). Publicizing Your Program: Website Evaluation, Design, and Marketing Strategies. *AACE Journal*, *15*(4), 437–471.

Senter, R. & Smith, E. A. (1967). *Automated readability index*. CINCINNATI UNIV OH.

SEO Glossary. (2018). Searchmetrics, Backlinks Definition.

Shang, W. (2018). A Comparison of the Historical Entries in Wikipedia and Baidu Baike. W *International Conference on Information* (s. 74–80). Springer.

Shen, A., Qi, J. & Baldwin, T. (2017). A Hybrid Model for Quality Assessment of Wikipedia Articles. W *Proceedings of the Australasian Language Technology Association Workshop 2017* (s. 43–52).

Shi, H. (2007). *Best-first decision tree learning* (Master's thesis, University of Waikato, Hamilton, NZ). COMP594.

Shvecov, A. V. (2011). Nekotorye metodicheskie podhody k jekonometricheskomu modelirovaniju vlijanija bjudzhetnoj politiki na jekonomiku. *Teorija i praktika obshhestvennogo razvitija*, (3).

Singh, P., Singh, P., Park, I. & Lee, J. (2009). Information sharing: a study of information attributes and their relative significance during catastrophic events. W *Cyber Security and Global Information Assurance: Threat Analysis and Response Solutions* (s. 283–305). IGI Global.

SISTRIX GmbH. (2018a). The secret of successful Websites.

SISTRIX GmbH. (2018b). What is the SISTRIX Visibility Index?

Soonthornphisaj, N. & Paengporn, P. (2017). Thai Wikipedia Article Quality Filtering Algorithm. W *Proceedings of the International MultiConference of Engineers and Computer Scientists* (Vol. 1).

Stefanowicz, B. (2010). *Informacja*. Warszawa: Szkoła Główna Handlowa Oficyna Wydawnicza.

Stein, K. & Hess, C. (2007). Does it matter who contributes: a study on featured articles in the german wikipedia. *HT '07: Proceedings of the eighteenth conference on Hypertext and hypermedia*, 171–174. doi:http://doi.acm.org/10.1145/1286240.1286290

Stvilia, B., Al-Faraj, A. & Yi, Y. J. (2009). Issues of cross-contextual information quality evaluation-The case of Arabic, English, and Korean Wikipedias. *Library and Information Science Research*, *31*(4), 232–239. doi:10.1016/j.lisr.2009.07.005

Stvilia, B., Gasser, L., Twidale, M. B. & Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American society for information science and technology*, *58*(12), 1720–1733.

Stvilia, B., Twidale, M. B., Smith, L. C. & Gasser, L. (2005a). Assessing information quality of a community-based encyclopedia. *Proc. ICIQ*, 442–454.

Stvilia, B., Twidale, M. B., Smith, L. C. & Gasser, L. (2005b). Assessing information quality of a community-based encyclopedia. *Proc. ICIQ*, 442–454.

Su, Q. & Liu, P. (2015). A Psycho-Lexical Approach to the Assessment of Information Quality on Wikipedia. W *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)* (Vol. 3, s. 184–187). doi:10.1109/WI-IAT.2015.23

Sumner, M., Frank, E. & Hall, M. (2005). Speeding up Logistic Model Tree Induction. W *9th European Conference on Principles and Practice of Knowledge Discovery in Databases* (s. 675–683). Springer.

Sundgren, B. (1973). *An infological approach to data bases*. National Central Bureau of Statistics, Sweden; University of Stockholm.

Suzuki, Y. & Yoshikawa, M. (2012). Mutual Evaluation of Editors and Texts for Assessing Quality of Wikipedia Articles. W *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration* (18:1–18:10). WikiSym '12. doi:10.1145/2462932.2462956

Swoboda, I. (2015). Jakość informacji. W S. Cisek i A. Januszko-Szakiel (Red.), *Zawód infobroker: polski rynek informacji* (s. 238–259). Wolters Kluwer.

Tacchini, E., Schultz, A. & Bizer, C. (2009). Experiments with wikipedia cross-language data fusion. W *Workshop on Scripting and Development*. Citeseer.

Teplitskiy, M., Lu, G. & Duede, E. (2017). Amplifying the impact of open access: Wikipedia and the diffusion of science. *Journal of the Association for Information Science and Technology*, *68*(9), 2116–2127.

The Network to Promote Linguistic Diversity (NPLD). (2018). Swedish.

Ustawa. (2008). O zasadach białoruskiej ortografii oraz interpunkcji. *Ustawa Republiki Białoruś*, (420-Z).

Velázquez, C. G., Cagnina, L. & Errecalde, M. L. (2017). On the feasibility of external factual support as wikipedia's quality metric.

Venables, W. N. & Ripley, B. D. (2002). Tree-based methods. W *Modern Applied Statistics with S* (s. 251–269). Springer.

Viera, A. J., Garrett, J. M. i in. (2005). Understanding interobserver agreement: the kappa statistic. *Fam Med*, *37*(5), 360–363.

Wakefield, J. (2017). Burger King advert sabotaged on Wikipedia.

Warncke-wang, M., Cosley, D. & Riedl, J. (2013). Tell Me More : An Actionable Quality Model for Wikipedia. W *WikiSym 2013* (s. 1–10). doi:10.1145/2491055.2491063

Warncke-Wang, M., Ranjan, V., Terveen, L. G. & Hecht, B. J. (2015). Misalignment Between Supply and Demand of Quality Content in Peer Production Communities. W *ICWSM* (s. 493–502).

Warrens, M. J. (2010). Inequalities between multi-rater kappas. *Advances in data analysis and classification*, *4*(4), 271–286.

Webb, G. (1999). Decision Tree Grafting From the All-Tests-But-One Partition, San Francisco, CA: Morgan Kaufmann.

Webster, J. & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Q. 26*(2), xiii–xxiii.

Węcel, K. & Lewoniewski, W. (2015). Modelling the Quality of Attributes in Wikipedia Infoboxes. W W. Abramowicz (Red.), *Business Information Systems Workshops* (Vol. 228, s. 308–320). Lecture Notes in Business Information Processing. doi:10.1007/978-3-319-26762-3_27

Wikimedia Strategic Planning. (nodate). Wikipedia Quality - Definition of quality.

Wikipedia - Featured article criteria. (nodate).

Wikipedia - Verifiability. (nodate).

Wikipedia Meta-Wiki. (2018a). List of Wikimedia projects by size.

Wikipedia Meta-Wiki. (2018b). List of Wikipedias.

WikiStats. (2018). List of largest Mediawikis.

Wilkinson, D. M. & Huberman, B. a. (2007a). Cooperation and quality in wikipedia. *Proceedings of the 2007 international symposium on Wikis WikiSym 07*, 157–164. doi:10.1145/1296951.1296968

Wilkinson, D. M. & Huberman, B. a. (2007b). Cooperation and quality in wikipedia. *Proceedings of the 2007 international symposium on Wikis WikiSym 07*, 157–164. doi:10.1145/1296951.1296968

Wöhner, T. & Peters, R. (2009). Assessing the quality of Wikipedia articles with lifecycle based metrics. *Proceedings of the 5th International Symposium on Wikis and Open Collaboration WikiSym 09*, 1. doi:10.1145/1641309.1641333

Wu, G., Harrigan, M. & Cunningham, P. (2011). Characterizing Wikipedia Pages Using Edit Network Motif Profiles. W *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents* (s. 45–52). SMUC '11. doi:10.1145/2065023.2065036

Wu, K., Zhu, Q., Zhao, Y. & Zheng, H. (2010). Mining the factors affecting the quality of Wikipedia articles. W *Information Science and Management Engineering (ISME), 2010 International Conference of* (Vol. 1, s. 343–346). IEEE.

Xu, H. (2015). What Are the Most Important Factors for Accounting Information Quality and Their Impact on AIS Data Quality Outcomes? *J. Data and Information Quality*, *5*(4), 14:1–14:22. doi:10.1145/2700833

Xu, H. & Koronios, A. (2005). Understanding information quality in e-business. *Journal of Computer Information Systems*, *45*(2), 73–82.

Xu, Y. & Luo, T. (2011). Measuring article quality in Wikipedia: Lexical clue model. *IEEE Symposium on Web Society*, (19), 141–146. doi:10.1109/SWS.2011.6101286

Yaari, E., Baruchson-Arbib, S. & Bar-Ilan, J. (2011). Information quality assessment of community generated content: A user study of Wikipedia. *Journal of Information Science*, *37*(5), 487–498.

Yu, L. (2011). *A developer's guide to the semantic Web*. Springer Science & Business Media.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J. & Auer, S. (2016). Quality assessment for linked data: A survey. *Semantic Web*, *7*(1), 63–93.

Zhang, S., Hu, Z., Zhang, C. & Yu, K. (2018). History-Based Article Quality Assessment on Wikipedia. W *Big Data and Smart Computing (BigComp), 2018 IEEE International Conference on* (s. 1–8). IEEE.

Spis tabel

Spis rysunków