STUDIA OECONOMICA POSNANIENSIA 2017, vol. 5, no. 12

DOI: 10.18559/SOEP.2017.12.7

Włodzimierz Lewoniewski, Krzysztof Węcel

Uniwersytet Ekonomiczny w Poznaniu, Wydział Informatyki i Gospodarki Elektronicznej, Katedra Informatyki Ekonomicznej

Autor do korespondencji: Włodzimierz Lewoniewski
wlodzimierz.lewoniewski@ue.poznan.pl

CECHY ARTYKUŁÓW ORAZ METODY ICH EKSTRAKCJI NA POTRZEBY OCENY JAKOŚCI INFORMACJI W WIKIPEDII

Streszczenie: Celem artykułu jest przedstawienie i klasyfikacja cech pozwalających na automatyczną ocenę jakości informacji zawartych w artykułach w Wikipedii. Na podstawie analizy literatury oraz własnych doświadczeń określono miary związane z artykułami, opisujące różne aspekty jakości. Dodatkowo zostały zaproponowane metody ekstrakcji cech służących do wyliczania miar jakości. Powiązania między artykułami w różnych językach stwarzają szanse w zakresie porównania i weryfikacji jakości informacji dostarczanych przez wikipedystów. Opracowany model może zatem znaleźć zastosowanie przy względnej ocenie jakości danych zawartych w strukturalnych częściach artykułów, tzw. infoboksach.

Słowa kluczowe: Wikipedia, DBpedia, jakość informacji, jakość danych, WikiRank.

Klasyfikacja JEL: C55, D8, L15, L86.

FEATURES OF WIKIPEDIA ARTICLES AND THEIR EXTRACTION METHODS FOR AUTOMATIC INFORMATION QUALITY ASSESSMENT

Abstract: This article presents and classifies features that can be extracted from Wikipedia articles for the purpose of automatic information quality assessment. Based on a state of the art analysis and our own experiments, specific measures for various aspects of quality have been defined. Additionally, an extraction method for various sources of features has been proposed. The links between articles in various languages offer op-

portunities for the comparison and verification of the quality of information delivered by wikipedians. The elaborated model can be used for the relative quality assessment of data contained in the structural parts of Wikipedia articles, namely info boxes.

Keywords: Wikipedia, DBpedia, quality of information, quality of data, WikiRank.

Wstęp

Obecnie w Wikipedii istnieje 285 aktywnych wersji językowych (https://en.wikipedia.org/wiki/List_of_Wikipedias). Największą jest wersja angielska, która posiada ponad 5,4 mln artykułów. W pierwszej dziesiątce największych edycji znajdują się również niemiecka, francuska, rosyjska i polska.

Ta internetowa encyklopedia stała się jednym z najważniejszych źródeł wiedzy na świecie. Każdy może przyczynić się do rozszerzenia jej zasobów. W maju 2017 roku liczba wyświetleń stron wynosiła około 516 mln dziennie dla wszystkich języków, a wersja polska była odwiedzana średnio 8,6 mln razy dziennie (https://stats.wikimedia.org/EN/TablesPageViewsMonthlyCombined.htm). W rankingu najpopularniejszych stron internetowych Wikipedia zajmuje piąte miejsce na świecie (http://www.alexa.com/topsites).

Codzienne zwiększa się liczba artykułów w każdym języku. Artykuły mogą być tworzone (edytowane) również przez anonimowych użytkowników. Autorzy nie muszą formalnie wykazywać swoich kwalifikacji z określonej dziedziny. Wikipedia nie ma centralnej redakcji czy grupy recenzentów, którzy mogliby kompleksowo podejść do weryfikacji wszystkich nowych i dotychczasowych artykułów. Te i inne problemy wywołały krytykę koncepcji Wikipedii, w szczególności wskazując na niską jakość informacji (https://pl.wikipedia.org/wiki/Krytyka_Wikipedii).

Zagadnienia jakości są jednak przedmiotem troski twórców Wikipedii. Praktycznie w każdej wersji językowej tej internetowej encyklopedii istnieje system wyróżnień dla artykułów wysokiej jakości. W angielskiej wersji Wikipedii najlepsze artykuły mają nazwę „Featured Article" (FA). Artykuły, które jeszcze nie spełniają wszystkich kryteriów FA, ale zbliżają się do ich jakości, mogą dostać również nieco niższe wyróżnienie „Good Article" (GA). W polskiej wersji odpowiednikami FA i GA są „Artykuł na medal" (ANM) i „Dobry artykuł" (DA). W celu otrzymania wyróżnienia artykuł musi być zgłoszony do nominacji przez użytkownika. W wyniku tego zostaje przeprowadzona dyskusja i odbywa się głosowanie, w którym każdy użytkownik może wyrazić zgodę lub nie na nadanie wyróżnienia dla konkretnego artykułu oraz wyjaśnić swój

punkt widzenia. Kryteria i zasady przyznawania wyróżnień w każdej wersji językowej mogą się zmieniać w czasie, co w konsekwencji może spowodować utratę wyróżnienia przez niektóre artykuły[1].

Oprócz wyróżnienia w niektórych wersjach językowych artykuł może otrzymać niższą ocenę. Taka pośrednia ocena może wskazywać na „dojrzałość" artykułu, czyli to, w jakim stopniu jest on zbliżony do wzorowego. W angielskiej wersji Wikipedii generalnie wyróżnia się siedem klas jakości artykułu (od najwyższej): FA, GA, A-class, B-class, C-class, Start, Stub. Warto zaznaczyć, że w odróżnieniu od wyższych (wzorowych) klas FA i GA inne (niższe) oceny nadawane są bez dyskusji społeczności i głosowania – każdy użytkownik może wystawić ocenę samodzielnie na podstawie przyjętych zasad. Niektóre wersje językowe stosują mniej rozwiniętą skalę ocen, np. w wersji polskiej oprócz wyróżnionych odpowiedników FA i GA istnieją również oceny (https://pl.wikipedia.org/wiki/Szablon:Stopnie_oceny_jako%C5%9Bci): Czwórka, Start, Zalążek (łącznie pięć klas).

W tabeli 1 pokazana jest liczebność artykułów w poszczególnych klasach jakości dla różnych języków. Z zestawienia wynika, że nie ma ogólnie przyjętego standardu klasyfikacji artykułów. Niektóre języki stosują rozwiniętą skalę ocen (en, ru), inne ograniczają się do dwóch, trzech 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 1 podobne klasy zostały pogrupowane). Z danej tabeli można wnioskować, że każda wersja językowa może mieć swój system klasyfikacji jakości artykułów, ale wszystkie stosują co najmniej dwie klasy wyróżnionych artykułów – odpowiedniki FA i GA. Jednak 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 w ogóle oceniona, np. w polskiej edycji udział artykułów nieocenionych stanowi ponad 99%.

Warto zaznaczyć, że w ramach jednego języka mogą być stosowane różne skale ocen dla poszczególnych projektów tematycznych. Na przykład w polskiej Wikipedii w projekcie „Gry komputerowe" stosuje się tylko trzy klasy jakości (ANM, DA, WER) (https://pl.wikipedia.org/wiki/Wikiprojekt:Gry_komputerowe/ Ocena_haseł), w projekcie „Filmy" można spotkać sześć różnych ocen (ANM, DA, PANDA, Czywiesz, Katalog, Problem) (https://pl.wikipedia.org/wiki/Wikiprojekt:Filmy/Ocena_haseł). W naszych badaniach skupiamy się na podstawowych skalach ocen jakości stosowanych w poszczególnych wersjach językowych.

[1] Dla polskiej wersji Wikipedii istnieje lista artykułów, które utraciły wyróżnienie: https://pl. wikipedia.org/wiki/Wikipedia:By%C5%82e_artyku%C5%82y_na_medal.

Tabela 1. Liczba artykułów w poszczególnych klasach jakości w różnych wersjach językowych Wikipedii (stan z kwietnia 2016 roku)

	Wersja językowa						
	be	**de**	**en**	**fr**	**pl**	**ru**	**uk**
Wszystkie artykuły	114 365	1 929 003	5 125 754	1 744 491	1 162 622	1 303 277	628 758
AMN (FA)	66	2 421	4 744	1 505	683	879	214
DA (GA)	109	3 837	23 921	2 515	2 107	2 38	578
Solidny artykuł						2 904	
A-class			795	1 538			
Czwórka					161		
Pełny artykuł						4 815	382
B-class			89 812	29 005			
Rozwinięty artykuł						17 999	1 951
C-class			214 895				
Artykuł w rozwoju						73 749	8 806
Start			1 285 974	244 839	1 255		
Zalążek (Stub)	575		2 567 378	779 485	1 655	92 682	6 652
Nieocenione	113 615	1 922 745	938 235	685 604	1 156 761	1 107 869	610 175

Każdy użytkownik może sprawdzić i poprawić jakość informacji zawartych w artykule na podstawie różnych kryteriów opisanych na specjalnych stronach w każdym języku[2], jednak wymaga to określonego nakładu czasowego. W literaturze naukowej można spotkać opracowania, w których zaproponowane są różne podejścia do automatycznej oceny jakości artykułów Wikipedii. Na podstawie różnych cech wysoko ocenionych (wyróżnionych) artykułów można próbować ocenić inne. Długość tekstu, liczba referencji, liczba obrazków i inne cechy artykułów mogą pomóc w ocenie jakości.

Celem prowadzonych przez nas badań jest zbudowanie modelu automatycznej oceny jakości artykułów Wikipedii, do czego można wykorzystać techniki eksploracji danych. Dokładność wyników w znacznej mierze zależy od doboru właściwych metod, a te często zależą od analizowanych danych. Wyniki analizy przy tym muszą być stosunkowo łatwe do zinterpretowania przez użytkownika.

[2] Dla polskiej wersji Wikipedii zasady napisania dobrego artykułu umieszczone są na stronie https://pl.wikipedia.org/wiki/Pomoc:Jak_napisać_doskonały_artykuł.

Większość badań dotyczących budowania modeli jakości artykułów Wikipedii skupia się na największej wersji językowej – angielskiej. W niniejszym opracowaniu bierzemy pod uwagę 7 języków: białoruski (be), niemiecki (de), angielski (en), francuski (fr), polski (pl), rosyjski (ru), ukraiński (uk). Pozwala to na budowanie modeli, które będą w stanie porównywać pomiędzy sobą jakość artykułów na jeden temat w różnych językach. Innymi słowy, celem naszych badań jest zaproponowanie kompleksowego modelu jakości informacji, który miałby zastosowanie do względnej oceny jakości.

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: 1) formalny proces oceny jakości artykułów Wikipedii, który dostarcza nam danych, 2) wielojęzyczność i powiązania między artykułami w różnych językach, stwarzające szanse w zakresie porównania i weryfikacji jakości wprowadzanych danych.

1. Przegląd literatury

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. 2005]. 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 i. in. 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 i in. 2009]. W przypadkach, gdy 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; Xu i Luo 2011]. Użytkownicy Wikipedii mogą umieszczać specjalne szablony w artykule, 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, Węcel i Abramowicz 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 [Wilkinson i Huberman 2007; Kittur i Kraut 2008]. Ważne jest to, aby w tej grupie redaktorów był chociaż jeden użytkownik z wysokim poziomem doświadczenia w edycji treści w Wikipedii [Arazy 2010]. 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 [Suzuki i Yoshikawa 2012; Halfaker, Kraut i Riedl 2009; Adler i De Alfaro 2007].

W niniejszej pracy zdecydowaliśmy się skupić przede wszystkim na tych aspektach, które mogą pomóc w poprawie jakości artykułu. Rozpatrujemy więc treść artykułu i jego metadane, a wpływ na listę współtwórców jest jednak ograniczony.

Podobnie jak w innych badaniach do budowania modelu może być stosowana dychotomiczna zmienna objaśniana [Xu i Luo 2011; Lex i in. 2012; Warncke-Wang, Cosley i Riedl 2013; Lewoniewski, Węcel i Abramowicz 2015] i jakość będzie modelowana jako prawdopodobieństwo przynależności do jednej z dwóch kategorii:

– k o m p l e t n e artykuły: klasy FA (ANM) i GA (DA),
– n i e k o m p l e t n e artykuły: wszystkie inne – rozwijające się (które należy dopracować) oraz nieocenione artykuły.

2. Model jakości informacji

Na podstawie literatury [Lih 2004; Stvilia i in. 2005; Hu i in. 2007; Wilkinson i Huberman 2007; Blumenstock 2008b; Blumenstock 2008a; Dalip i in. 2009; Lipka i Stein 2010; Dalip i in. 2011; Anderka 2013; Warncke-Wang, Cosley i Riedl 2013] oraz własnych badań [Lewoniewski, Węcel i Abramowicz 2015; Węcel i Lewoniewski 2015; Lewoniewski, Węcel i Abramowicz 2016] niżej zostały wymienione cechy, które mogą być brane pod uwagę przy budowaniu modeli oceny jakości artykułów Wikipedii. Modele te uwzględniają wielość źródeł, z których można pozyskać miary do oceny jakości informacji. Do modelu zostały wybrane cechy różnego rodzaju. Dla łatwiejszego zrozumienia zostały one podzielone na kategorie: statystyki tekstowe, części mowy, wzory czytelności, podobieństwo słów, struktura artykułu i inne.

2.1. Statystyki tekstowe

W tej kategorii znajdują się cechy, które pochodzą z tekstu artykułu i dotyczą głównie liczby znaków i słów. Ilekroć jest mowa o wskaźniku, chodzi o udział procentowy określonego zjawiska, np. wskaźnik pytań wyznaczamy poprzez podzielenie liczby pytań przez liczbę zdań. Wyjaśnienia może wymagać również pojęcie szumu – są to wszelkie znaki, które nie niosą treści, ale są wykorzystywane np. do formatowania tekstu czy tworzenia szablonów.

1. Liczba znaków
2. Długość w bajtach
3. Liczba liter
4. Liczba trigramów*
5. Wskaźnik złożoności słów
6. Wskaźnik informacja/szum
7. Wskaźnik długich zdań
8. Wskaźnik długich słów
9. Najdłuższe zdanie
10. Liczba 1-sylabowych słów
11. Wskaźnik 1-sylabowych słów
12. Liczba paragrafów
13. Długość paragrafu
14. Liczba pytań
15. Wskaźnik pytań
16. Liczba zdań
17. Długość zdania
18. Wskaźnik krótkich zdań
19. Długość najkrótszego zdania
20. Liczba sylab
21. Liczba słów
22. Długość słów
23. Długość sylab

* Trigram – ciąg występujących po sobie trzech liter.

2.2. Części mowy

Podobnie jak w poprzedniej kategorii analizowane są parametry dotyczące treści artykułu, jednak tutaj mamy do czynienia z bardziej złożoną analizą, która wymaga dodatkowych predefiniowanych zasobów (słowników) dla każdej wersji językowej.

1. Wskaźnik czasowników pomocniczych
2. Wskaźnik spójników
3. Wskaźnik zdań ze spójnikiem
4. Wskaźnik zdań z zaimkiem pytającym
5. Wskaźnik nominalizowanych słów
6. Wskaźnik biernych zdań
7. Wskaźnik przyimków
8. Wskaźnik zdań z przyimkiem
9. Wskaźnik zdań z zaimkiem
10. Wskaźnik zaimków
11. Wskaźnik zdań ze spójnikiem podrzędnym
12. Wskaźnik czasowników „być" („to be")

2.3. Wzory czytelności

Wymienione niżej wzory pozwalają ocenić złożoność tekstu. Niestety większość dostępnych cech dotyczy języka angielskiego, a opracowanie ich dla innych języków nie jest trywialne.

1. Automatyczny wskaźnik czytelności
2. Bormuth Index
3. Coleman-Liau Index
4. FORCAST Readability
5. Flesch Reading Ease
6. Flesch-Kincaid
7. Gunning Fog Index
8. Läsbarhetsindex
9. Miyazaki EFL Readability Index
10. New Dale-Chall
11. SMOG Grading

2.4. Charakterystyka słów

Analizowane są poszczególne słowa w tekście. Tutaj również dla przeprowadzenia analizy niezbędne są specjalne predefiniowane słowniki dla poszczególnych języków.

1. Wskaźnik słów powszechnych
2. Wskaźnik skomplikowanych słów powszechnych
3. Wskaźnik Peacock*
4. Wskaźnik stop-słów
5. Wskaźnik prostych słów

* Peacock (z ang. paw) – oznacza używanie słów ozdobników, nieniosących znaczenia.

2.5. Struktura artykułu

W tej kategorii wymienione są cechy dotyczące organizacji i struktury artykułu. Do ich utworzenia wykorzystuje się specjalne wiki-znaczniki, np. dla referencji to jest `<ref>...</ref>`, dla szablonów {{...}}. Niektóre cechy można podzielić na podgrupy, np. przy liczeniu referencji można brać pod uwagę każde odwołanie do dowolnego źródła lub liczyć tylko unikatowe referencje. W przypadku szablonów wskazujących na luki jakości należy brać pod uwagę niespójność w ich definiowaniu pomiędzy różnymi wersjami językowymi. Należy zaznaczyć, że nie wszystkie cechy da się wyznaczyć, np. z powodu braku określonej struktury.

1. Szum (liczba znaków)
2. Liczba liter bez szumu
3. Informatywność
4. Kompletność
5. Liczba kategorii
6. Liczba plików
7. Liczba nagłówków
8. Liczba obrazków
9. Liczba obrazków na sekcję
10. Liczba infoboksów
11. Długość abstraktu
12. Wskaźnik abstraktu
13. Liczba referencji
14. Liczba sekcji z referencjami
15. Liczba referencji na sekcję
16. Liczba referencji na długość tekstu
17. Liczba sekcji
18. Długość sekcji
19. Odchylenie długości sekcji
20. Zagnieżdżanie w sekcji
21. Długość najkrótszej sekcji
22. Długość najkrótszej podsekcji
23. Liczba podsekcji
24. Długość podsekcji
25. Zagnieżdżanie w podsekcji
26. Liczba podsekcji 2 poziomu
27. Długość podsekcji 2 poziomu
28. Długość największej sekcji
29. Długość największej podsekcji
30. Długość największej podsekcji 2 poziomu
31. Liczba tabeli
32. Liczba szablonów
33. Liczba sekcji z ciekawostkami
34. Liczba szablonów o lukach w jakości

2.6. Historia edycji

W tej kategorii wskazujemy cechy związane z rozwojem artykułu w czasie.

1. Wskaźnik edycji przez aktywnych użytkowników
2. Wskaźnik edycji przez administratorów
3. Wiek artykułu
4. Wiek na edycję
5. Wskaźnik edycji przez anonimowych użytkowników
6. Łączność
7. Obieg
8. Liczba edycji
9. Wskaźnik obiegu edycji
10. Liczba edycji na autora
11. Odchylenie liczby edycji na autora
12. Liczba autorów
13. Wskaźnik autorów
14. Wskaźnik modyfikacji linii tekstu
15. Wskaźnik przypadkowych autorów
16. Wskaźnik jakości autorów
17. Wskaźnik przywróconych edycji
18. Wskaźnik zarejestrowanych użytkowników
19. Liczba przywróconych edycji
20. Czas przywrócenia edycji
21. Stopień zabezpieczenia edycji

2.7. Parametry sieciowe

Dane parametry pokazują, jak dany artykuł jest związany z innymi artykułami Wikipedii.

1. Liczba kategorii
2. Liczba linków wewnętrznych
3. Liczba linków wewnętrznych na długość tekstu
4. Liczba słabych linków wewnętrznych
5. Współczynnik grupowania
6. Liczba linków zewnętrznych
7. Liczba linków zewnętrznych na sekcję
8. Liczba wersji językowych artykułu
9. PageRank
10. Współczynnik wzajemności

2.8. Popularność artykułu

Częstość czytania artykułu również może mieć wpływ na jakość artykułu. Można założyć, że im większa liczba użytkowników czyta artykuł, tym szybciej zostaną wychwycone ewentualne błędy i częściej będą wprowadzone tam zmiany.

1. Liczba obserwujących użytkowników
2. Liczba odwiedzin za 90 dni
3. Liczba odwiedzin za 30 dni
4. Mediana liczby odwiedzin za 90 dni
5. Mediana liczby odwiedzin za 30 dni
6. Średnia liczby odwiedzin za 30 niepustych dni
7. Pozycja artykułu w rankingu za rok

2.9. Charakterystyka dyskusji

Każdemu artykułowi towarzyszy oddzielna strona dyskusji (https://pl.wikipedia.org/wiki/Pomoc:Strona_dyskusji), która służy do wymiany opinii, zgłaszania uwag i rozwiązywania konfliktów związanych z treścią artykułu.

1. Długość strony dyskusji
2. Liczba edycji na stronie dyskusji
3. Liczba autorów na stronie dyskusji
4. Wiek strony dyskusji
5. Liczba nagłówków na stronie dyskusji
6. Liczba szablonów na stronie dyskusji
7. Liczba linków wewnętrznych na stronie dyskusji
8. Liczba linków zewnętrznych na stronie dyskusji
9. Liczba słów na stronie dyskusji
10. Liczba zdań na stronie dyskusji

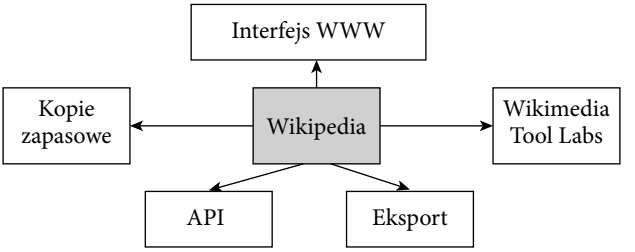
2.10. Podsumowanie

Przedstawione powyżej cechy jakości wykorzystywane są do wyliczania miar jakości na podstawie budowanych modeli eksploracji danych (*data mining*). Zagadnienie to leży jednak poza zakresem niniejszego artykułu.

3. Metody pozyskiwania cech do automatycznej oceny jakości informacji

W związku z wielością źródeł, z których można pozyskać cechy 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 1 pokazane są różne podejścia do pozyskiwania danych dotyczących artykułów Wikipedii. Najprostszy jest interfejs WWW. Kolejnym stosunkowo prostym narzędziem dostępu do artykułów Wikipedii jest specjalny formularz (https://pl.wikipedia.org/wiki/Specjalna:Eksport) pozwalający eksportować jeden lub wiele artykułów w formacie XML wraz z historią edycji (do 1000 edycji).

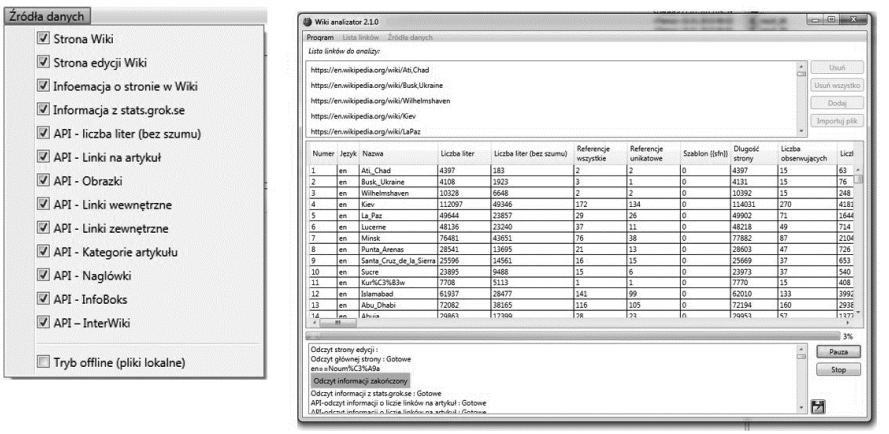


Rysunek 1. Możliwości dostępu do danych artykułów Wikipedii

Co miesiąc tworzona jest kompletna kopia wszystkich artykułów Wikipedii w poszczególnych wersjach językowych w postaci wikitekstu (kodu źródłowego) i metadanych w różnych formatach (w tym XML) oraz surowych baz danych w postaci SQL. Wszystkie te pliki są swobodnie dostępne na specjalnym serwerze (https://dumps.wikimedia.org). Wykorzystanie kopii zapasowych do analizy dużej liczby artykułów wygląda interesująco, jednak w odróżnieniu od innych metod wymaga większego nakładu czasu do pozyskania danych dotyczących konkretnych artykułów. 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 70 GB.

Wikipedia Tool Labs (http://tools.wmflabs.org) (dawniej Toolserver) jest środowiskiem uruchomieniowym dla użytkowników Wikipedii do tworzenia różnego oprogramowania, które jest skierowane na ułatwienia korzystania z tej encyklopedii. Ponad 1000 narzędzi pozwala m.in. otrzymać różnego rodzaju statystyki dotyczące artykułów. Jednak te programy 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 w momencie 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 2 pokazany jest interfejs graficzny tego programu oraz źródła, z których pobiera on dane.



Rysunek 2. Interfejs graficzny programu WikiAnalizator wraz z wykazem źródeł danych

Serwis API działa dla każdego języka i dostępny jest pod adresem określonym według szablonu: https://{język}.wikipedia.org/w/api.php?action={ustawienia}, gdzie {język} to skrót wersji językowej, {ustawienia} – ustawienia zapytania[3]. Na przykład w celu otrzymania określonych informacji dotyczących artykułu „Polska" w polskim języku w formacie JSON należy wykorzystać następujące wywołania:

– listę wszystkich nagłówków:
  https://pl.wikipedia.org/w/api.php?action=parse&format=json&prop=links&page=Polska
– listę wszystkich szablonów:
  https:///pl.wikipedia.org/w/api.php?action=parse&format=json&prop=templates&page=Polska
– listę wszystkich obrazków:
  https://pl.wikipedia.org/w/api.php?action=parse&format=json&prop=images&page=Polska.

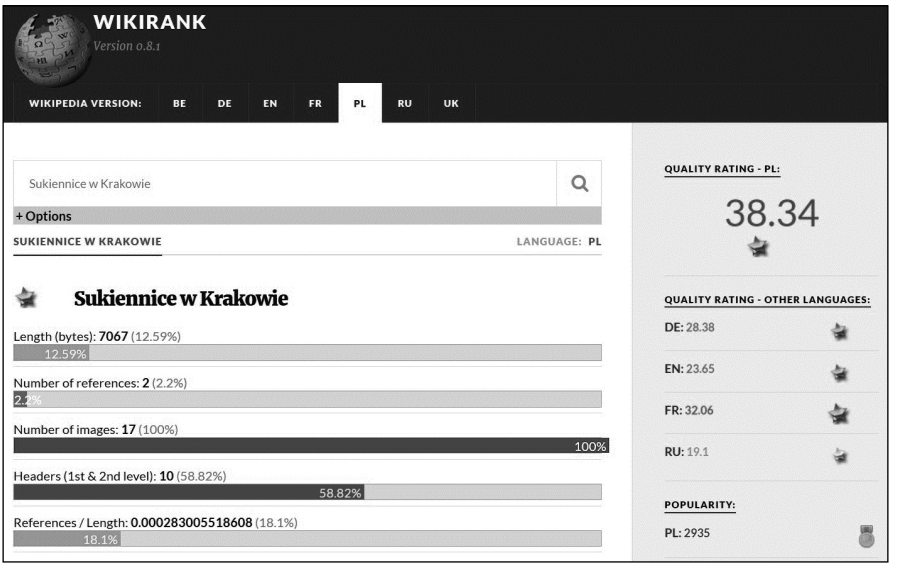
Podsumowanie

Z wykorzystaniem części z wymienionych w niniejszej pracy cech (ponad 80) zostały przeprowadzone wstępne próby tworzenia modeli jakości przy użyciu technik eksploracji danych. Ich wyniki wskazują na istotne różnice między modelami dla poszczególnych języków – w sumie 7 wersji językowych Wikipedii: angielskiej, niemieckiej, francuskiej, rosyjskiej, polskiej, ukraińskiej i białoruskiej. Niniejszy artykuł skupia się przede wszystkim na cechach, dlatego przy analizie modeli wzięto pod uwagę istotność poszczególnych zmiennych dla objaśnienia zmienności w jakości artykułów. Na przykład dla 5 z 7 badanych języków bardzo istotny jest parametr „liczba liter bez szumu", a jedną ze wspólnych najistotniejszych cech w pozostałych 2 wersjach językowych jest „liczba referencji". Są również takie cechy, które są statystycznie istotne tylko dla jednego języka: np. średnia odwiedzin za 30 dni, liczba unikatowych autorów. Większe podobieństwo modeli można obserwować przy analizie artykułów o tej samej tematyce (np. pisarze, gry komputerowe, filmy) w różnych językach. Interesujące jest to, że nawet w ramach jednej wersji językowej Wikipedii modele różnią się w zależności od tematyki artykułu.

Proponowane modele mogą pomóc w poprawie jakości artykułów Wikipedii przez identyfikację najlepszej wersji językowej konkretnego artykułu.

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

Niektóre algorytmy eksploracji danych pozwalają określić istotność parametrów w modelach jakości, które później mogą być wykorzystane do porównania artykułów w różnych językach. Tę właściwość wykorzystaliśmy przy tworzeniu serwisu Wikirank, który służy do obliczania tzw. względnej jakości artykułów (http://wikirank.net). Na rysunku 3 pokazany jest przykład oceny artykułu „Sukiennice w Krakowie" w 7 wersjach językowych Wikipedii na podstawie 5 zmiennych: długość artykułu, liczba referencji, liczba nagłówków 1 i 2 poziomu, liczba obrazków oraz stosunek referencji do długości artykułu. W skali od 0 do 100 najwyższą ocenę uzyskała wersja polska (38,34 pkt), co może oznaczać najlepszą jakość w tym języku spośród 7 badanych.



Rysunek 3. Analiza artykułu „Sukiennice w Krakowie" w polskiej Wikipedii w projekcie WikiRank.net

Źródło: [http://wikirank.net/pl/Sukiennice_w_Krakowie].

Nasze badania mogą pomóc w poprawie jakości danych w DBpedii (http://dbpedia.org), jednej z najbardziej znanych semantycznych baz danych, która jest wzbogacana poprzez ekstrahowanie faktów z artykułów różnych wersji językowych Wikipedii. Nasze modele w połączeniu z innymi wskaźnikami mogą pomóc w identyfikacji danych lepszej jakości [Węcel i Lewoniewski 2015].

Bibliografia

Adler, B.T., De Alfaro, L., 2007, *A Content-driven Reputation System for the Wikipedia*, in: *Proceedings of the 16th International Conference on World Wide Web WWW 07 7.Generic*, s. 261–270.

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

Arazy, O., 2010, *Determinants of Wikipedia Quality: the Roles of Global and Local Contribution Inequality*, in: *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, s. 233–236.

Blumenstock, J.E., 2008a, *Automatically Assessing the Quality of Wikipedia Articles*, Raport techniczny.

Blumenstock, J.E., 2008b, *Size Matters: Word Count as a Measure of Quality on Wikipedia*, in: *Proceedings of the 17th International Conference on World Wide Web*, s. 1095–1096.

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*, in: *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, s. 295–304.

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, vol. 2, iss. 3, s. 1–30.

Halfaker, A., Kraut, R., Riedl, J., 2009, A Jury of Your Peers: Quality, Experience and Ownership in Wikipedia, in: *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, s. 1–10.

Hu, M., Lim, E.-P., Sun, A., Lauw, H.W., Vuong, B.-Q., 2007, *Measuring Article Quality in Wikipedia*, in: *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, s. 243–252.

Kittur, A., Kraut, R.E., 2008, *Harnessing the Wisdom of Crowds in Wikipedia*, in: *Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work*, s. 37–46.

Lewoniewski, W., Węcel, K., Abramowicz, W., 2015, *Analiza porównawcza modeli jakości informacji w narodowych wersjach Wikipedii*, w: Porębska-Miąc, T. (red.), *Systemy Wspomagania Organizacji SWO*, Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, Katowice, s. 133–154.

Lewoniewski, W., Węcel, K., Abramowicz, W., 2016, *Quality and Importance of Wikipedia Articles in Different Languages*, w: Dregvaite, G., Damescevicius, R. (ed.), *Information and Software Technologies. ICIST 2016*, Communications in Computer and Information Science, no. 639, s. 613–624.

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,*

in: *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, s. 7–10.

Lih, A., 2004, *Wikipedia as Participatory Journalism: Reliable Sources? Metrics for Evaluating Collaborative Media as a News Resource*, in: *5th International Symposium on Online Journalism*, s. 31.

Lipka, N., Stein, B., 2010, *Identifying Featured Articles in Wikipedia: Writing Style Matters*, in: *Proceedings of the 19th International Conference on World Wide Web*, s. 1147–1148.

Stein, K., Hess, C., 2007, *Does It Matter Who Contributes: A Study on Featured Articles in the German Wikipedia*, in: *HT '07: Proceedings of the Eighteenth Conference on Hypertext and Hypermedia*, s. 171–174.

Stvilia, B., Twidale, M.B., Smith, L.C., Gasser, L., 2005, *Assessing Information Quality of a Community-Based Encyclopedia*, in: *Proceedings of the 2005 International Conference on Information Quality*, s. 442–454.

Suzuki, Y, Yoshikawa, M., 2012, *Mutual Evaluation of Editors and Texts for Assessing Quality of Wikipedia Articles*, in: *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, Linz, s. 18:1–18:10.

Warncke-Wang, M., Cosley, D., Riedl, J., 2013, *Tell Me More: An Actionable Quality Model for Wikipedia*, in: *Proceedings of the 9th International Symposium on Open Collaboration*, s. 1–10.

Węcel, K., Lewoniewski, W., 2015, *Modelling the Quality of Attributes in Wikipedia Infoboxes*, in: Abramowicz, W. (ed.), *Business Information Systems Workshops*, Lecture Notes in Business Information Processing, vol. 228, s. 308–320.

Wilkinson, D.M., Huberman, B.A., 2007, *Cooperation and Quality in Wikipedia*, in: *Proceedings of the 2007 International Symposium on Wikis*, s. 157–164.

Wöhner, T., Peters, R., 2009, *Assessing the Quality of Wikipedia Articles with Lifecycle Based Metrics*, in: *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, s. 1–10.

Xu, Y., Luo, T., 2011, *Measuring Article Quality in Wikipedia: Lexical Clue Model*, in: *IEEE Symposium on Web Society 19*, s. 141–146.