

Włodzimierz Lewoniewski

Uniwersytet Ekonomiczny w Poznaniu
Wydział Informatyki i Gospodarki Elektronicznej
wlodzimierz.lewoniewski@ue.poznan.pl

Budowanie modeli jakości artykułów na określony temat w narodowych wersjach Wikipedii

Streszczenie: Wikipedia, największa encyklopedia internetowa, jest bogatym zasobem informacji. Jej wykorzystanie w działalności gospodarczej jest jednak ograniczone ze względu na trudności z oceną jakości informacji. Wychodząc naprzeciw oczekiwaniom konsumentów informacji, społeczność Wikipedii opracowała system oceny jakości artykułów, określający, czy informacje w artykułach są pełne, wiarygodne i aktualne. Ponieważ każdy artykuł w tej encyklopedii zostaje oceniony tylko w wyniku inicjatywy użytkownika, pojawia się problem dużej liczby nieocenionych artykułów. Poza tym treść ocenionego artykułu może zostać później poprawiona, co może się wiązać z koniecznością dokonania powtórnej oceny. Biorąc pod uwagę fakt, że w każdej wersji językowej Wikipedii istnieją swoiste kryteria oceny artykułów oraz dodatkowo artykuły z niektórych dziedzin mają własne uszczegółowione kryteria należy stwierdzić, że kwestia ręcznej weryfikacji jakości umieszczonej informacji w tej popularnej encyklopedii staje się poważnym wyzwaniem. W niniejszej pracy została przeprowadzona analiza porównawcza miar jakości dla artykułów na różne tematy w różnych wersjach językowych Wikipedii oraz zostały zbudowane modele, które mogą w przyszłości posłużyć do oceny jakości artykułów w poszczególnych językach z uwzględnieniem tematyki informacji. Dodatkowo pokazano różnice w istotności miar w modelach jakości w zależności od wersji językowej oraz tematów artykułów.

Słowa kluczowe: Wikipedia, encyklopedia, Internet, jakość informacji, ocena, DBpedia, WikiRank

Klasyfikacja JEL: C55, D8, L15, L86

Wstęp

Wikipedia – bezpłatna wielojęzyczna encyklopedia internetowa, która jest tworzona, utrzymywana i kontrolowana przez wolontariuszy z całego świata. Wikipedia nie ma centralnej redakcji lub redaktora naczelnego, dlatego każdy zainteresowany może współtworzyć treść.

Obecnie Wikipedia ma około 300 wersji językowych (Meta-Wiki, 2017a), z których najbardziej rozwinięte są: angielska, niemiecka, francuska, rosyjska, włoska, hiszpańska, polska. Każda z tych wersji zawiera ponad milion artykułów na różnorodne tematy, największa z nich (angielska) – prawie 5 mln. Liczba artykułów ciągle rośnie.

Każdy artykuł w Wikipedii może posiadać linki dające możliwość szybkiego przejścia do stron na ten sam temat w innych wersjach językowych (tzw. interwiki). Wówczas każda wersja językowa tego samego artykułu może powstawać niezależnie, co może prowadzić do niezgodności informacji pomiędzy tymi wersjami. Może więc pojawić się pytanie: w jakiej wersji językowej informacja o konkretnym podmiocie jest bardziej kompletna, aktualna, dokładna i wiarygodna. Inaczej mówiąc, dla użytkownika znającego dwa i więcej języków byłoby korzystne wiedzieć, na którą wersję językową artykułu należy zwrócić większą uwagę.

Użytkownicy edytujący treści w Wikipedii nie muszą formalnie wykazywać swoich kwalifikacji w określonej dziedzinie, co znacznie rozszerza grono potencjalnych autorów i powoduje szybszą rozbudowę informacji w danej encyklopedii. Jednak stwarza to również zagrożenie obniżenia jakości umieszczonych informacji. W Wikipedii istnieje system weryfikacji zmian wprowadzonych przez nowych lub mniej rzetelnych autorów do artykułu. Takie zmiany mogą być weryfikowane przez bardziej kompetentnych użytkowników – redaktorów. Proces sprawdzenia takiej informacji może trwać od kilkunastu sekund do ponad roku w zależności od tematu artykułu i rozwinięcia danej wersji językowej encyklopedii.

Poważnym wyzwaniem jest kwestia weryfikacji jakości ciągle tworzonych i edytowanych tekstów Wikipedii. Jakość artykułów może być oceniana na podstawie różnych ich cech, na przykład długości, referencji, liczby ilustracji, liczby linków. Miary te można wykorzystać do automatycznej oceny jakości artykułu. Jednak w każdej wersji językowej Wikipedii mogą one mieć swoją wagę – w jednej bardziej istotne są referencje, w innej objętość tekstu. Celem badawczym tego opracowania jest analiza najistotniejszych cech artykułów, które mogą być wykorzystane do automatycznej ich klasyfikacji na różne tematy według stopnia dojrzałości z punktu widzenia określonych narodowych wersji Wikipedii: białoruskiej (BE), niemieckiej (DE), angielskiej (EN), polskiej (PL), rosyjskiej (RU), ukraińskiej (UK).

Wyróżnione artykuły w Wikipedii

W Wikipedii istnieje system wyróżnień dla artykułów, które uważane są za wzorowe. Takie artykuły istnieją praktycznie w każdej wersji językowej Wikipedii. Muszą one spełniać określone kryteria jakości, takie jak: dokładność, neutralność, kompletność, styl napisania i inne. W angielskiej wersji Wikipedii najlepsze artykuły nazywają się Featured Article (FA). Artykuły, które nie spełniają wszystkich kryteriów FA, ale zbliżają się do ich jakości, mogą dostać inną wysoką ocenę – Good Article (GA). W polskiej wersji odpowiednikami FA i GA są „artykuł na medal” i „dobry artykuł”. Nadawanie tych wyróżnień następuje w wyniku

dyskusji i głosowania, w których każdy użytkownik może się zgodzić lub nie z nadaniem artykulowi wyższej oceny jakości oraz wyjaśnić swój punkt widzenia. Zanim nastąpi taka dyskusja, ktoś z użytkowników musi zgłosić artykuł do oceny.

W niektórych wersjach językowych, oprócz wyróżnienia, artykuł może dostać inne oceny jakości. Taka ocena może pokazać, na ile „dojrzały” jest artykuł oraz w jakim stopniu jest on zbliżony do wzorcowego. Na przykład w największej wersji językowej Wikipedii (angielskiej) generalnie wyróżnia się 7 klas jakości artykułu (od najwyższej): Featured Article (FA), Good Article (GA), A-class, B-class, C-class, Start, Stub (Wikipedia, 2017). Oceny poniżej FA i GA nadawane są bez dyskusji i głosowania – każdy użytkownik może wystawić ocenę samodzielnie na podstawie przyjętych w danej wersji językowej kryteriów.

Każda wersja językowa może mieć swój system klasyfikacji i nie wszystkie stosują rozwiniętą skalę ocen, jednak zazwyczaj każda ma dwie najwyższe oceny – odpowiedniki FA i GA.

Wystawiane przez użytkowników oceny mogą być nieobiektywne i niezgodne z zasadami. Ocena może być również nieaktualna – treść artykułu zmienia się w czasie i może on wymagać ponownego oceniania. Dotyczy to głównie niższych klas jakości, w których przypadku użytkownik może samodzielnie dokonywać oceny artykułu.

Zasady przyznawania poszczególnych ocen w każdej wersji językowej mogą się zmieniać w czasie i nawet artykuły z wyróżnieniem mogą je stracić. Na przykład w angielskiej Wikipedii od momentu powstania najwyższej klasy jakości FA standardy jakości wielokrotnie się zmieniały, co spowodowało utratę wyróżnienia przez ponad 1000 artykułów (Wikipedia, 2017b), które dostały je wcześniej na mniej „surowych” zasadach. W polskiej wersji ponad 200 artykułów straciło „medal” (Wikipedia, 2017c).

Ocena każdego artykułu wymaga inicjatywy użytkowników, w związku z tym pojawia się problem dużej liczby nieocenionych tekstów. W tabeli 1 pokazano

Tabela 1. Liczba artykułów o różnej jakości w różnych wersjach językowych (na czerwiec 2015 r.). Podobne klasy jakości są pogrupowane

Nazwa/Język	BE	DE	EN	PL	RU	UK
Liczba artykułów	89 923	1 828 090	4 741 168	1 099 441	1 198 199	557 590
Featured Article	54	2 383	4 481	648	768	204
Good Article	97	3 785	21 697	1 993	2 041	531
Solid article					1 661	
A-class			821			
Czwórka				157		
Pełny artykuł					3 978	195
B-class			80 919			
Rozwinięty artykuł					15 492	707
C-class			194 063			
Artykuł w rozwoju					60 801	3 490
Start			1 177 495	928		
Stub (załączek)	860		2 634 800	2 071	429 926	189 742

Źródło: opracowanie własne.

system klasyfikacji artykułów w 6 językach (BE, DE, EN, PL, RU, UK) oraz liczba artykułów należących do poszczególnych klas jakości.

Według stanu na czerwiec 2015 r., w angielskiej wersji Wikipedii nieocenionych artykułów jest 13,2% liczby wszystkich artykułów w tym języku. W mniej rozwiniętych wersjach językowych udział takich artykułów jest znacznie większy: białoruska – 98,9%, niemiecka – 99,7%, polska – 99,5%, rosyjska – 57,1%, ukraińska – 65,1%. Zatem można stwierdzić, że w Wikipedii istnieje duża liczba artykułów nieposiadających żadnej oceny.

Jakość artykułu a przyrost treści

Ocena jakości artykułów Wikipedii w literaturze naukowej jest stosunkowo rozwiniętym tematem. Zaproponowano wiele modeli, które automatycznie klasyfikują artykuły Wikipedii pod względem jakości na podstawie miar dotyczących artykułów. Badania można podzielić na trzy grupy – ocena parametrów ilościowych artykułów (tj. liczba liter), ocena treści, analiza zachowania twórców treści. Każda z tych metod umożliwia z różną precyzją automatyczne klasyfikowanie artykułów w ramach angielskiej wersji Wikipedii.

Stosunkowo prostym podejściem jest pomiar objętości treści (długość tekstu, liczba referencji, nagłówków, obrazków, edycji i inne). Badania pokazują, że zazwyczaj artykuły bardziej dojrzałe (wyższej jakości) zawierają więcej treści. Na przykład długość artykułu jest doskonałym wskaźnikiem tego, czy artykuł jest dobry czy nie (Blumenstock, 2008). Wyróżnione artykuły są dość długie, wykorzystują w spójny sposób referencje, edytowane przez setki redaktorów i mają tysiące wersji (Hu i in., 2007; Blumenstock, 2008; Wöhner, Peters, 2009).

Bardziej złożone badania skupiają się wokół analizy jakościowej treści artykułu. W jednej z prac został wykorzystany tzw. indeks czytelności FOG, który określa stopień przystępności tekstu (Hasan Dalip i in., 2009). W przypadkach, kiedy objętość treści w artykułach jest podobna, lepszy artykuł będzie mieć więcej informacji faktycznej (Lex i in., 2012). Styl napisania i różnorodność wykorzystanych słów również wpływa na jakość artykułu (Xu, Luo, 2011; Lipka, Stein, 2010).

W odniesieniu do analizy zachowania użytkowników istnieją podejścia, które oceniają ich doświadczenie i reputację. Artykuły wysokiej jakości mają dużą liczbę edycji i dużą liczbę redaktorów, którzy charakteryzują się wysokim poziomem współpracy (Wilkinson, Huberman, 2007; Kittur, Kraut, 2008). Ważne jest to, żeby w tej grupie redaktorów był chociażby jeden użytkownik z wysokim poziomem doświadczenia w edycji treści w Wikipedii (Arazy, 2010). Szczególne znaczenie ma reputacja użytkownika, który dokonał pierwszej edycji artykułu (Stein, Hess, 2007). Może być liczona na podstawie „przetrwania” tekstu, który on umieścił (Adler, De Alfaro, 2007; Halfaker i in., 2009; Suzuki, Yoshikawa, 2012).

Zachowanie użytkowników może pomóc w procesie automatyzacji przypisania artykułu do odpowiedniej klasy, jednak nie w jego poprawie. Innymi słowy, jeżeli chcemy dowiedzieć się, jakich treści brakuje w artykule dla lepszej oceny, musimy analizować odpowiednie cechy tego artykułu.

Ponieważ analiza jakościowa treści artykułów w różnych językach wymaga posiadania specjalistycznych słowników, w tej pracy dla oceny dojrzałości artykułu będziemy wykorzystywać różne jego parametry ilościowe. Do listy takich parametrów zostały dobrane już znane z literatury naukowej oraz proponowane nowe w celu odnalezienia najefektywniejszych z nich.

Warto zaznaczyć, że w większości badań dotyczących automatycznej oceny jakości artykułów w Wikipedii budowanie i weryfikacja modeli dokonywana była na podstawie angielskiej wersji Wikipedii bez uwzględnienia zakresu informacji.

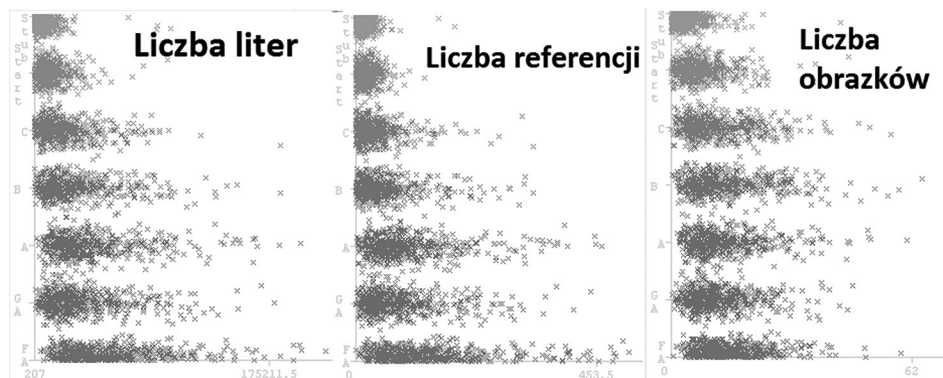
Lista atrybutów

Na podstawie literatury oraz własnych eksperymentów stworzono listę atrybutów, która będzie brana pod uwagę w dalszej analizie. Techniczne skróty tych atrybutów przedstawione są poniżej:

A1 – Autorzy	A13 – LinkiWiZ	A25 – Obrazki P
A2 – Długość	A14 – LinkiZ	A26 – Obserwujące
A3 – Edycje	A15 – LOG (długość)	A27 – Referencje
A4 – Edycje/Autorzy	A16 – LOG (liczba liter)	A28 – Referencje/długość
A5 – Informatywność 1	A17 – Nagłówek 1+2	A29 – Referencje/Liczba Liter
A6 – Informatywność 2	A18 – Nagłówek 1+2+3	A30 – Ref/Liczba Liter bez Szumu 1
A7 – Języki	A19 – Nagłówki 1	A31 – Ref/Liczba Liter bez Szumu 2
A8 – Liczba Liter	A20 – Nagłówki 2	A32 – Referencje/Linki WiZ
A9 – Liczba Liter bez Szumu 1	A21 – Nagłówki 3	A33 – Referencje U
A10 – Liczba Liter bez Szumu 2	A22 – Obrazki	A34 – Szablony
A11 – Linki NA	A23 – Obrazki/długość	A35 – Szum 1
A12 – LinkiW	A24 – Obrazki NP	A36 – Szum 2

Niżej przedstawiono opis niektórych z atrybutów:

- Autorzy – liczba unikatowych autorów, uczestniczących w tworzeniu (edycji) artykułu,
- Długość – długość artykułu w bajtach,
- Edycje – ile razy artykuł został edytowany od momentu jego utworzenia,
- Języki – liczba wersji językowych artykułu w Wikipedii,
- LinkiNA – liczba linków na artykuł,
- LinkiW – liczba linków wewnętrznych do innych artykułów w ramach Wikipedii (tzw. wikilinks),
- LinkiZ – liczba linków do zewnętrznych stron internetowych,
- LinkiWiZ – suma LinkiW i LinkiZ,
- ObrazkiP – liczba obrazków bezpośrednio powiązanych z artykułem,
- ObrazkiNP – liczba obrazków niepowiązanych bezpośrednio z artykułem (np. znaczki flag państw w tekście, techniczne obrazki z szablonów itd.),
- Obrazki – liczba wszystkich obrazków w artykule, inaczej to ObrazkiP + ObrazkiNP,



Ryc. 1. Rozkład atrybutów (od lewej – liczba liter, liczba referencji, liczba obrazków) w poszczególnych klasach jakości angielskiej Wikipedii (najśłabsza klasa jakości Stub znajduje się u góry, najlepszej klasa FA – na dole)

Źródło: opracowanie własne w programie WEKA.

- Szum – wszystko, co nie dotyczy bezpośrednio tekstu (czyli techniczne symbole, formatowanie, szablony itp.), różnica pomiędzy Szum1 i Szum2 – polega na sposobie liczenia tego szumu,
- Nagłówki 1 – liczba nagłówków pierwszego poziomu,
- Nagłówki 2 – liczba nagłówków drugiego poziomu,
- Nagłówki 1+2+3 – suma liczby nagłówków 1, 2 i 3 poziomu w artykule,
- ReferencjeU – liczba referencji unikatowych wykorzystanych w artykule,
- Referencje – liczba wszystkich referencji, z powtórzeniami,
- Nagłówki – suma Nagłówki 1 i Nagłówki 2,
- Informatywność – liczona według wzoru: $0,6 * \text{Szum} + 0,3 * \text{Obrazki}$.

Niektóre atrybuty artykułu są pozytywnie skorelowane z jego jakością, tzn. przy większej jakości artykułu mamy większą wartość atrybutu. W angielskiej wersji językowej przykładem takiego atrybutu może być liczba liter (patrz ryc. 1).

Zbiór danych

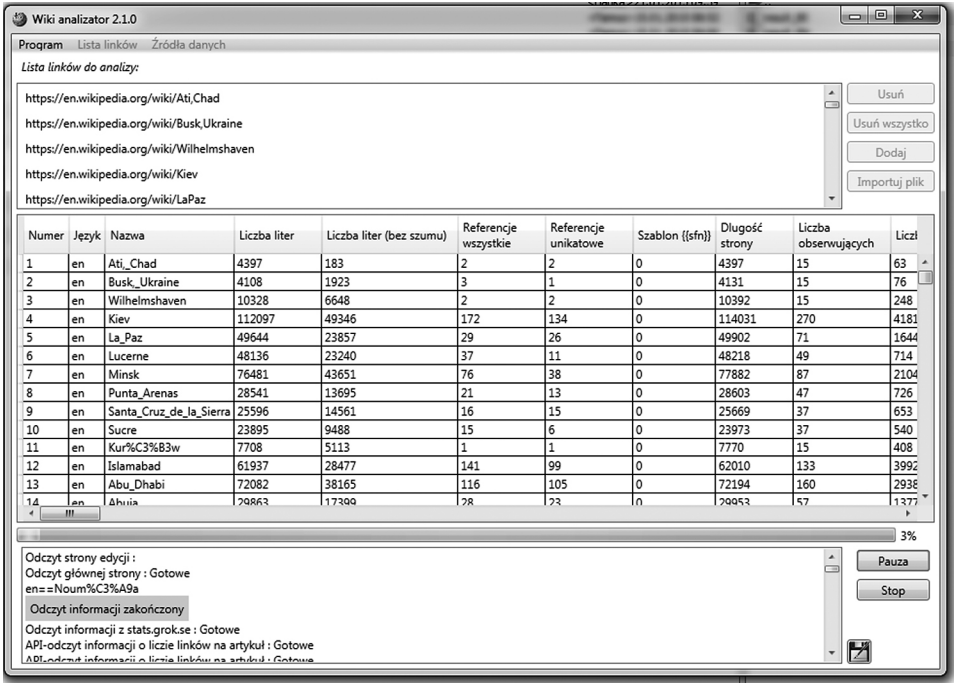
Analiza będzie przeprowadzona oddzielnie dla artykułów związanych z filmami, grami komputerowymi i pisarzami w każdej z 6 wersji językowych Wikipedii: białoruskiej (BE), niemieckiej (DE), angielskiej (EN), polskiej (PL), rosyjskiej (RU), ukraińskiej (UK). O przynależności artykułu do określonego tematu mówi kategoria, która została do niego przypisana przez użytkowników encyklopedii. Zestawienie z liczbą artykułów z poszczególnych kategorii w 6 językach przedstawiono w tabeli 2.

Dla artykułów z każdej kategorii w poszczególnych językach zostały wyeks-
trahowane metadane, tj. liczba liter, liczba edycji, liczba autorów, liczba obraz-
ków, liczba linków, liczba referencji, liczba nagłówków i inne. Do wykonania tego

Tabela 2. Liczba artykułów w poszczególnych kategoriach w 6 wersjach językowych Wikipedii (stan na wrzesień 2015 r.)

Temat	Język					
	BE	DE	EN	PL	RU	UK
Filmy	176	30369	114472	21269	28366	11041
Gry komputerowe	94	3102	22745	2765	5406	1141
Pisarze	1971	60424	99913	15191	19786	8798

Źródło: opracowanie własne.



Ryc. 2. Wiki Analizator – program do pobierania metadanych dotyczących artykułów Wikipedii (proces ekstrakcji metadanych)

Źródło: opracowanie własne.

zadania wykorzystano własny program Wiki Analizator (część interfejsu użytkownika przedstawiono na ryc. 2).

Analiza zgromadzonych danych

Dobór metody

Wcześniejsze badania pokazały efektywność wykorzystywania drzew klasyfikacyjnych w zadaniach automatycznej identyfikacji klasy jakości artykułów na

podstawie ich atrybutów ilościowych (Xu, Luo, 2011; Warncke-wang i in., 2013; Lewoniewski i in., 2015). Jedną z zalet drzew klasyfikacyjnych jest fakt, że wygenerowany model można łatwo przedstawić w postaci zbioru reguł (Morzy, 2006). Jednym z częstych zastosowań drzew jest analiza istotności (ważności) atrybutów, która może pomóc w odnalezieniu najważniejszych cech artykułu wpływających na jego klasę jakości.

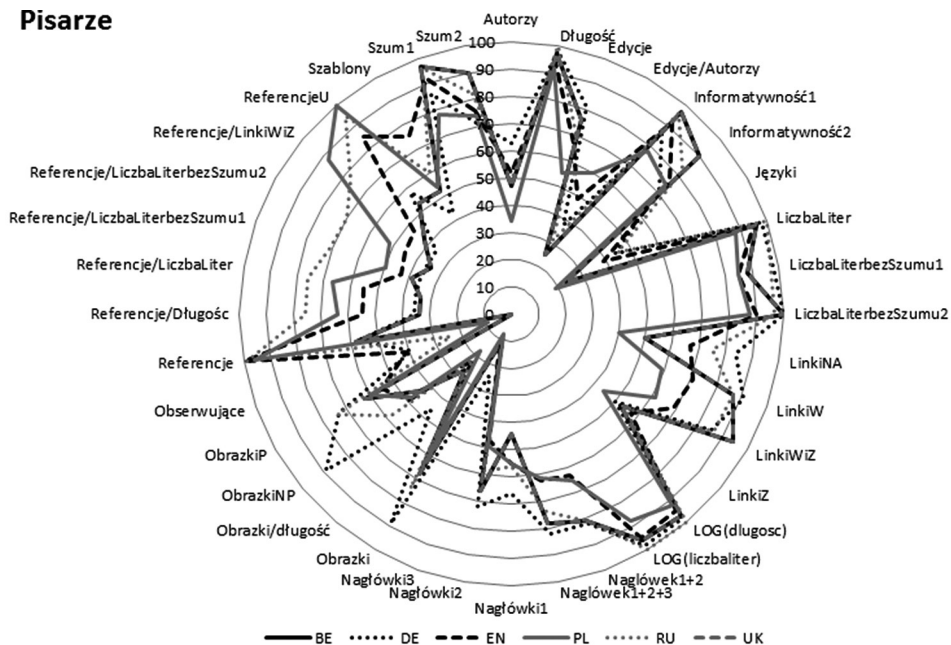
W celu doboru odpowiedniej metody zweryfikowano 5 modeli klasyfikacyjnych: C&RT, Boosted Trees, Random Forest, SVM, Sieci Neuronowe. Wybrany został model C&RT, który miał najwyższy poziom precyzji na zebranych danych.

Zmienna celu jako zmienna binarna

Jak wspomniano, w każdym języku może występować mniej lub bardziej rozwinięty system klasyfikacji jakości artykułów, jednak istnieją wspólne dla wszystkich 6 języków 2 najwyższe klasy jakości – FA oraz GA. Zatem możemy podzielić wszystkie artykuły na kompletne (w systemie binarnym – 1) i te, które należy dopracować (w systemie binarnym – 0). Do kompletnych będą należały artykuły z najwyższych klas odpowiednich angielskim FA i GA, a artykuły z niższych klas przechodzą do grupy „należy dopracować”.

Przy takim podejściu jakość może być modelowana jako prawdopodobieństwo przynależności do jednej z klas. Za stosowaniem takiego binarnego podziału

Pisarze



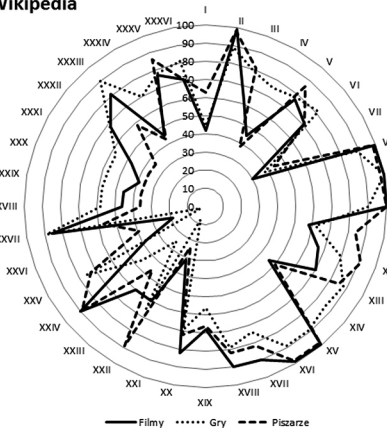
Ryc. 3. Istotność atrybutów w metodzie C&RT artykułów o pisarzach w różnych językach Wikipedii przy binarnej zmiennej zależnej

Źródło: opracowanie własne.

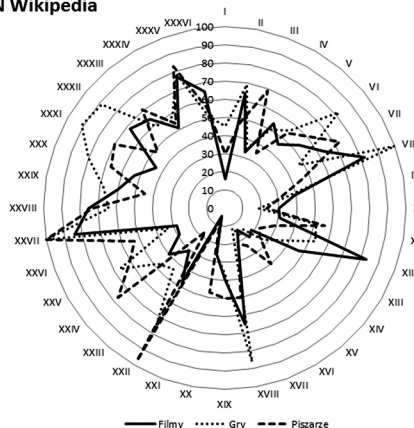
przemawia również fakt, że klasy FA i GA (należące do „kompletnych”) są bardziej wiarygodne na tle innych (niższych) klas jakości, m.in. poprzez większą kontrolę przez społeczność Wikipedii. Poza tym w literaturze naukowej już opisana została efektywność takiego podziału (Lex i in., 2012; Warncke-wang i in., 2013).

Do analizy każdego tematu w każdej wersji językowej grupy kompletnych i niekompletnych artykułów muszą mieć podobną liczebność. Na przykład, jeżeli w polskiej wersji Wikipedii o pisarzach napisano 15 191 art., w tym do najwyższych klas FA należy 18 artykułów i do klasy GA – 31 artykułów (razem w grupie „kompletne” – 49 artykułów), to do grupy „należy dopracować” dobierzemy losowo również 49 artykułów z tej samej kategorii.

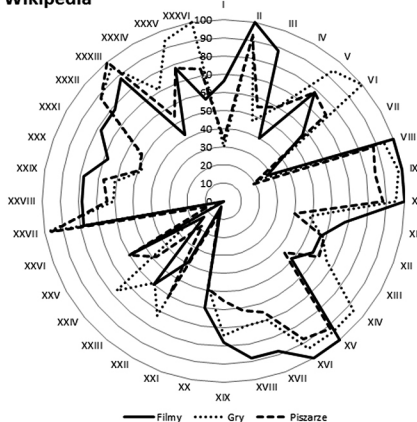
DE Wikipedia



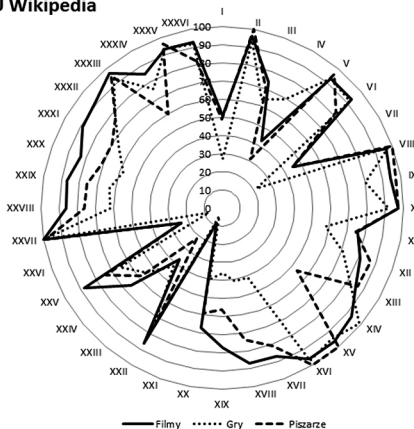
EN Wikipedia



PL Wikipedia



RU Wikipedia



Ryc. 4. Istotność atrybutów w metodzie C&RT artykułów o pisarzach, filmach, grach komputerowych w ramach angielskiej, polskiej, rosyjskiej Wikipedii przy binarnej zmiennej zależnej (skrótów zgodnie z rozdz. 4.2)

Źródło: opracowanie własne.

Obliczenie istotności atrybutów

Istotność w C&RT mierzona jest w skali od 0 do 100. Im większa wartość, tym większa istotność atrybutu. Na rycinie 3 pokazane są istotności atrybutów z kategorii „Pisarze” w 6 wersjach językowych Wikipedii. Jak widać, modele jakości artykułów w tej samej kategorii są zbliżone w różnych językach. Podobne zestawienia przygotowano dla kategorii „Filmy” oraz „Gry komputerowe”. Natomiast inna sytuacja powstaje przy porównaniu istotności atrybutów artykułów o różnej tematyce w ramach jednego języka. Na rycinie 4 pokazane są istotności zmiennych artykułów w każdej z 3 kategorii w niemieckiej, angielskiej, polskiej oraz rosyjskiej Wikipedii. Podobne analizy przeprowadzono dla artykułów białoruskiej i ukraińskiej Wikipedii (w celu oszczędzenia miejsca pracy została przedstawiona tylko część wykresów; pełny raport zob. Lewoniewski.info, 2015). Poziom precyzji modelu C&RT wyniósł około 95–97% w zależności od wersji językowej i kategorii artykułów.

Wnioski

Na podstawie wykresów można stwierdzić, że modele jakości artykułów o tej samej tematyce w różnych językach są podobne. Jeżeli rozpatrywać wykresy z analizy istotności atrybutów artykułów o różnej tematyce w ramach jednego języka – mamy do czynienia z większymi różnicami w modelach jakości.

Przy porównaniu modeli jakości pomiędzy wersjami językowymi widoczna jest różnica w wybranych istotnych miarach do określenia jakości artykułów. Dodatkowo uwidoczniona jest różnica pomiędzy modelami określającymi jakości artykułów na dany temat w ramach jednej wersji językowej. Na przykład w ramach polskojęzycznej Wikipedii dla określenia jakości artykułów o grach komputerowych najistotniejszymi są informatywność oraz liczba nagłówków 1–3 poziomu. Z drugiej strony, dla artykułów opisujących filmy w tej samej wersji językowej ważniejsze są miary dotyczące długości artykułu.

Można zauważyć, że istotność niektórych atrybutów ma wspólny charakter dla wszystkich kategorii w ramach jednego języka. Na przykład – liczba wszystkich referencji dla angielskiej Wikipedii i prawie podobna istotność tego atrybutu w rosyjskiej Wikipedii. Jednak w polskiej Wikipedii atrybut ten ma różną istotność: dla artykułów o filmach ma on niższe znaczenie niż dla artykułów o grach lub pisarzach. Innymi słowy częściej można spotkać rozbieżności w istotności w ramach danej wersji językowej w różnych kategoriach. W mniejszym stopniu dotyczy to rozbieżności istotności atrybutów w ramach jednej kategorii, ale w różnych językach.

W związku z powyższymi wynikami należy stwierdzić, że poszczególne wersje językowe Wikipedii różnią się pod względem ważności atrybutów w ocenie jakości w zależności od tematu i wersji językowej. Stosunkowo wysoki poziom precyzji (95–97%) stał się możliwy dzięki zastosowaniu binarnej klasyfikacji artykułów. Zatem mamy do czynienia z bardziej wyodrębnionymi istotnymi zmiennymi,

które mogą w sposób bardziej precyzyjny określić stopień dojrzałości artykułu. Związane jest to przede wszystkim z niezgodnością w standardach klasyfikacji dojrzałości artykułu pomiędzy różnymi wersjami językowymi Wikipedii. Dodatkowym atutem jest fakt, że dla grupy kompletnych artykułów istnieje mniejsze zagrożenie niezgodnej lub błędnej oceny – jest to konsekwencja większej kontroli ze strony społeczności Wikipedii nad przydzieleniem artykułu do najwyższych klas jakości w porównaniu do innych klas.

W analizie istniały również ograniczenia. Zauważono, że w nielicznych sytuacjach artykuł może być przypisany od razu do dwóch, trzech „konkurencyjnych” kategorii – np. „Gry komputerowe” i „Filmy”. Zazwyczaj dotyczy to artykułów z grupy jakościowej „należy dopracować”, w związku z tym dany artykuł można było bez problemu zastąpić innym artykułem z próby losowej. W przyszłości może to być poprawione bardziej złożoną analizą tematyki artykułu: np. będą też brane pod uwagę wykorzystane w artykule dodatkowe elementy (tj. infoboksy, szablony) lub struktura tekstu.

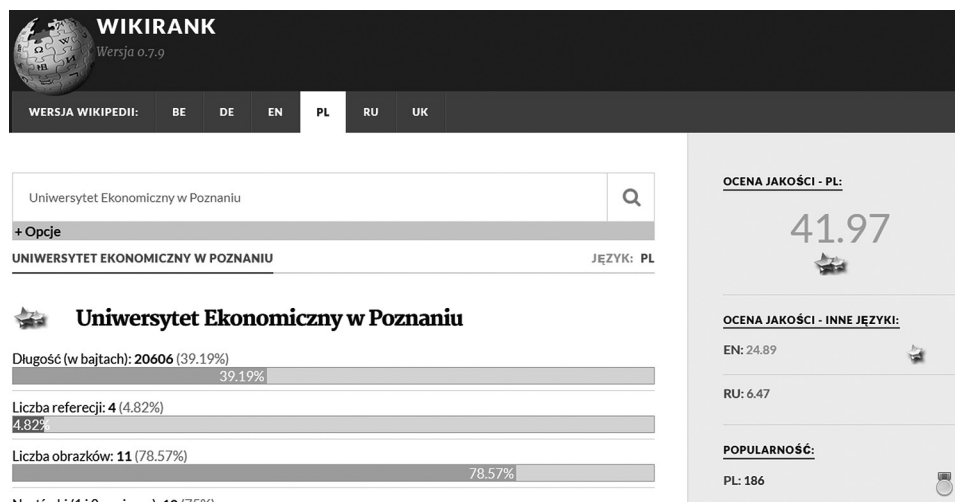
W związku z powyższym można stwierdzić, że dla większej precyzji w automatycznej ocenie jakości należy brać pod uwagę nie tylko wersję językową Wikipedii, w której jest napisany artykuł, ale również jego tematykę.

Możliwe zastosowania i dalsze prace

Znalezienie istotnych atrybutów może pomóc w budowaniu modeli automatycznej oceny jakości artykułów Wikipedii w różnych językach oraz umożliwi porównywanie tego samego artykułu w różnych językach. Jednym z przykładów przyszłego rozwoju modelu jest projekt WikiRank.net, który obecnie porównuje jakość artykułów Wikipedii w 6 językach, nadając każdej wersji odpowiednią liczbę punktów w skali 0–100. Na rycinie 5 przedstawiony jest przykład działania tego projektu przy analizie artykułu *Uniwersytet Ekonomiczny w Poznaniu* (WikiRank, 2017). Innymi słowy, rozwijanie i stosowanie metody nie tylko pozwoli na bardziej precyzyjne automatyczne ocenianie jakości artykułu w ramach poszczególnych wersji językowych, ale również da możliwość względnej oceny artykułów o tym samym podmiocie w różnych językach. Może to być przydatne dla użytkowników Wikipedii, którzy znają dwa lub więcej języków. Niektórzy z nich wykorzystują bardziej rozwinięte wersje językowe Wikipedii i przenoszą informację do innych języków. Model może im pomóc wybrać tę wersję językową, w której zawarte informacje mają większą jakość.

W przyszłości analiza modeli w różnych podobnych językach może pozwolić na bardziej precyzyjne zidentyfikowanie modeli oceny jakości dla mniej rozwiniętych wersji językowych, takich jak np. białoruska.

W niniejszej pracy została wykorzystana zmienna celu jako zmienna binarna, jednak dla zwiększenia szczegółowości analizy można rozważyć rozbudowanie tego systemu klasyfikacji, jednocześnie zachowując jego uniwersalność do wszystkich (lub większości) wersji językowych Wikipedii. Jednym z pierwszych kroków może być podzielenie grupy „kompletnych”, w której znajdują się klasy



Ryc. 5. Analiza artykułu *Uniwersytet Ekonomiczny w Poznaniu* w polskiej Wikipedii w projekcie WikiRank.net

FA i GA. Związane jest to przede wszystkim z tym, że każda wersja językowa ma oddzielne klasy oraz istnieją w nich różnice pomiędzy standardami jakości. Obydwie te klasy wskazują na wysoką ocenę artykułów, jednak klasa FA jest wyższa niż GA. Dlatego w przyszłych badaniach można budować modele jakości na podstawie 3 grup artykułów: FA, GA i „należy dopracować”.

Jednym z interesujących kierunków rozwoju przedstawionego podejścia jest wykorzystanie zmiennej celu ze skalą porządkową lub interwałową. Jednak wstępne analizy pokazują, że znalezienie funkcji odwzorowania klas jakości na konkretne wartości na skali interwałowej jest dużym wyzwaniem.

Zwiększenie precyzji modelu może być osiągnięte również poprzez wprowadzenie do niego zmiennych jakościowych, takich jak np. jakość referencji podanych w artykule. Większa jakość źródeł podanych w postaci referencji może wpływać na jakość informacji w artykule. Rozważane jest również badanie struktury treści i styl napisania artykułu.

W dalszych badaniach przedstawiony model może pomóc w odnalezieniu zależności pomiędzy jakością całego artykułu a jakością poszczególnych faktów (atrybutów) umieszczonych w infoboksie danego artykułu. Te fakty są ekstrahowane z różnych wersji językowych Wikipedii do jednej z najbardziej znanych semantycznych baz danych – projektu DBpedia. Badania pokazują, że mimo swojej popularności i przydatności, baza ta może zawierać nieścisłości oraz nieprawdziwe i niepełne informacje (Zaveri i in., 2016). Proponowany model może pomóc w poprawie jakości danych, poprzez identyfikację tych wersji językowych, na podstawie których można będzie uzupełniać fakty w innych językach DBpedii (Węcel, Lewoniewski, 2015).

Obiecujący jest również kierunek analizy wielojęzycznych użytkowników Wikipedii. Takie osoby jednocześnie dokonują zmiany w różnych wersjach językowych

i zazwyczaj skupiają się na określonych tematach, przenosząc podobną informację o podmiocie z jednego języka do innego. Dodatkowo tacy użytkownicy mogą się skupiać na dodawaniu określonego typu treści (np. obrazki, referencje itd.). Analiza taka pozwoli m.in. określić ich wpływ na ocenę jakości w poszczególnych wersjach językowych.

Proponowany sposób analizy i budowania modeli jakości może być wykorzystany nie tylko w Wikipedii. Oprogramowanie MediaWiki, na podstawie której działa Wikipedia, umożliwia wszystkim chętnym tworzenie własnych encyklopedii na własnych serwerach. Niektóre z takich wolnych encyklopedii mają ponad 12 mln stron (Meta-Wiki, 2017b). Są również inne serwisy działające na zasadach wolnej encyklopedii (tj. Wikia.com), w których podobne modele jakości mogą również być skuteczne.

Literatura

- 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 (Generic), s. 261.
- Arazy O. (2010). *Determinants of Wikipedia Quality: the Roles of Global and Local Contribution Inequality*. New York, s. 233–236.
- Blumenstock J.E. (2008). *Size matters: word count as a measure of quality on wikipedia*. Proceedings of the 17th International Conference on World Wide Web, ACM, s. 1095–1096.
- Halfaker A., Kraut r., Riedl J. (2009). *A Jury of Your Peers: Quality, Experience and Ownership in Wikipedia*. WikiSym'09, s. 1–10.
- Hasan Dalip D., André Gonçalves M., Cristo M., Calado P. (2009). *Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia*. Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, ACM, s. 295–304.
- Hu M., Lim E.P., Sun A., Lauw H.W., Vuong B.Q. (2007). *Measuring article quality in wikipedia: models and evaluation*. Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, ACM, s. 243–252.
- 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, s. 37.
- Lewoniewski W., Węcel K., Abramowicz W. (2015). *Analiza porównawcza modeli jakości informacji w narodowych wersjach Wikipedii*. *Prace Naukowe/Uniwersytet Ekonomiczny w Katowicach*, s. 133–154.
- Lewoniewski.info (2015). *Budowanie modeli jakości artykułów na określony temat w narodowych wersjach Wikipedii* (<http://lewoniewski.info/files/wiki2015gts.pdf>; dostęp: 2.05.2017).
- Lex E., Voelske M., Errecalde M., Ferretti E., Cagnina L., Horn C., Granitzer M. (2012). *Measuring the quality of web content using factual information*. Proceedings of the 2nd joint WICOW/AIRWeb workshop on web quality, ACM, s. 7–10.
- Lipka N., Stein B. (2010). *Identifying Featured Articles in Wikipedia: Writing Style Matters*. Proceedings of the 19th International Conference on World Wide Web (2010), s. 1147–1148.
- Meta-Wiki (2017a). *List of Wikipedias* (https://meta.wikimedia.org/wiki/List_of_Wikipedias; dostęp: 2.05.2017).
- Meta-Wiki (2017b). *List of largest wikis* (https://meta.wikimedia.org/wiki/List_of_largest_wikis; dostęp: 2.05.2017).
- Morzy M. (2006). *Eksploracja danych – przegląd dostępnych metod i dziedzin zastosowań*. VI edycja Hurtownie Danych i Business Intelligence. Warszawa: Centrum Promocji Informatyki, 11.
- 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, s. 171–174.

- Suzuki Y., Yoshikawa M. (2012). *Mutual Evaluation of Editors and Texts for Assessing Quality of Wikipedia Articles*. Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration, ACM, s. 1–10.
- Warncke-wang M., Cosley D., Riedl J. (2013). *Tell Me More: An Actionable Quality Model for Wikipedia*. WikiSym, s. 1–10.
- Węcel K., Lewoniewski W. (2015). *Modelling the quality of attributes in Wikipedia infoboxes*. International Conference on Business Information Systems. Springer International Publishing, s. 308–320.
- Wilkinson D.M., Huberman B.A. (2007). *Cooperation and quality in wikipedia*. Proceedings of the 2007 International Symposium on Wikis WikiSym 07, s. 157–164.
- Wikipedia (2017a). *Wikipedia: WikiProject assessment* (http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment; dostęp: 2.05.2017).
- Wikipedia (2017b). *Wikipedia: Former featured articles* (http://en.wikipedia.org/wiki/Wikipedia:Former_featured_articles; dostęp: 2.05.2017).
- Wikipedia (2017c). *Wikipedia: Byłe artykuły na medal* (http://pl.wikipedia.org/wiki/Wikipedia:Byłe_artykuły_na_medal; dostęp: 02.05.2017).
- WikiRank (2017). *Uniwersytet Ekonomiczny w Poznaniu* (http://wikirank.net/pl/Uniwersytet_Ekonomiczny_w_Poznaniu; dostęp: 2.05.2017).
- 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, s. 1.
- Xu Y., Luo T. (2011). *Measuring article quality in Wikipedia: Lexical clue model*. IEEE Symposium on Web Society, 19, 141–146.
- 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.

Building articles quality models on a defined topic in national versions of Wikipedia

Abstract: Wikipedia, the largest online encyclopedia, is a rich resource of information. However, its use in business is limited because of the difficulty in assessing the quality of information contained in that encyclopedia. In response to consumer expectations, the Wikipedia community has developed a quality assessment system that determines whether the information contained in the articles is complete, reliable and up-to-date. Because every article in this encyclopedia evaluated solely on the initiative of the user, there is a problem of a large number of unassessed articles. In addition, the content of the assessed article can be revised later, which may require a re-evaluation. In each language version of Wikipedia there are specific criteria for articles assessment, and additionally articles from some areas have their own detailed criteria. Thus, the issue of manual verification of the quality of information placed in this popular encyclopedia becomes a serious challenge. In this paper, a comparative analysis of quality measures for articles on various topics in different language versions of Wikipedia was performed, and an attempt was made to build models that can be used to assess the articles quality in particular language, taking into account the topics. In addition, differences in the significance of measures in quality models depending on the language version and the topic of the articles have been shown.

Keywords: Wikipedia, Encyclopedia, Internet, information quality, assessment, DBpedia, WikiRank

JEL codes: C55, D8, L15, L86