re-BUS

You have downloaded a document from
RE-BUŚ
repository of the University of Silesia in Katowice

Title: Miary podobieństw łańcuchów znakowych a deduplikacja rekordów w bibliograficznych bazach danych

Author: Anna Małgorzata Kamińska

Citation style: Kamińska Anna Małgorzata. (2017). Miary podobieństw łańcuchów znakowych a deduplikacja rekordów w bibliograficznych bazach danych. "Przegląd Biblioteczny" (2017), z. 4, s. 477-495.

Uznanie autorstwa - Użycie niekomercyjne - Bez utworów zależnych Polska - Licencja ta zezwala na rozpowszechnianie, przedstawianie i wykonywanie utworu jedynie w celach niekomercyjnych oraz pod warunkiem zachowania go w oryginalnej postaci (nie tworzenia utworów zależnych).

UNIWERSYTET ŚLĄSKI W KATOWICACH
Biblioteka Uniwersytetu Śląskiego
Ministerstwo Nauki i Szkolnictwa Wyższego

PRZEGLĄD BIBLIOTECZNY 2017 z. 4
PL ISSN 0033-202X

ANNA MAŁGORZATA KAMIŃSKA
Uniwersytet Śląski w Katowicach
Instytut Bibliotekoznawstwa i Informacji Naukowej
e-mail: anna.kaminska@us.edu.pl

MIARY PODOBIEŃSTW ŁAŃCUCHÓW ZNAKOWYCH A DEDUPLIKACJA REKORDÓW W BIBLIOGRAFICZNYCH BAZACH DANYCH

Anna Małgorzata Kamińska, dr, adiunkt w Instytucie Bibliotekoznawstwa i Informacji Naukowej Uniwersytetu Śląskiego w Katowicach, pracownik Biblioteki Głównej Politechniki Śląskiej w Gliwicach. W 2016 r. obroniła rozprawę doktorską „Informacja naukowa o górnictwie w świetle wydawnictw ciągłych uczelni technicznych w Polsce (1945-1989)" na Wydziale Filologicznym Uniwersytetu Śląskiego w Katowicach. Jej zainteresowania naukowe skupiają się wokół trzech uzupełniających się obszarów: informatologia, graficzne języki komunikacji oraz wizualizacja informacji.

SŁOWA KLUCZOWE: Bibliograficzne bazy danych. Deduplikacja rekordów. Podobieństwo łańcuchów znakowych. Scalanie rekordów.

ABSTRAKT: **Teza/cel artykułu** – Celem artykułu jest przedstawienie metody deduplikacji/łączenia (ang. *deduplication/linkage*) rekordów opisujących jednostki bibliograficzne w bazach danych opartej na miarach podobieństw łańcuchów znakowych. Algorytm opracowano na podstawie własnych doświadczeń nabytych podczas tworzenia bibliograficznej bazy danych oraz podczas realizacji badań bibliometrycznych, na podstawie publicznie dostępnych bibliograficznych baz danych. Formalny opis metody zilustrowano przykładami zaczerpniętymi z krajowej bibliograficznej bazy CYTBIN. **Metody badawcze** – Opracowanie metody wymagało przeglądu architektur informacyjnych wybranych krajowych bibliograficznych baz danych, określenia typologii problemów ich dotykających, wynikających nie tylko z przyjętych modeli składowania danych, ale i budowy graficznych interfejsów użytkownika, którymi są zasilane, analizy i wyboru miar podobieństw łańcuchów znakowych oraz ostatecznie zaproponowania miary złożonej umożliwiającej ewaluację podobieństwa rekordów bibliograficznych w oparciu o wartości ich atrybutów składowych.

Wyniki – Przedstawione na przykładzie danych pochodzących z wybranej bazy bibliograficznej wyniki pozwoliły empirycznie zweryfikować użyteczność zaproponowanej metody. Dodatkowo dokonano analizy rozkładu podobieństwa rekordów bibliograficznych bazy CYTBIN określanego na podstawie zaproponowanej metody złożonej i metody opartej na mierze Jaro-Winkler wyliczanej dla tytułów jednostek bibliograficznych. **Wnioski** – Zaproponowana metoda, po dostrojeniu jej parametrów do specyfiki (występujących anomalii) konkretnych baz bibliograficznych, może być wprost zastosowana do poprawy jakości opisów bibliograficznych w nich gromadzonych, zarówno w proaktywnym modelu pracy (przed zatwierdzeniem opisu przez operatora), jak i modelu reaktywnym (weryfikacja wszystkich lub nowo zgromadzonych rekordów wykonywana np. w czasie mniejszego obciążenia systemu w dobowych odstępach czasu).

WSTĘP

Wszelkiego rodzaju przekazom informacji towarzyszy nieodłączne ryzyko utraty ważnych jej części, bądź zniekształcenia jej pierwotnej treści. Systemy języków naturalnych, które ukształtowały się i rozwinęły na drodze ewolucji, wytworzyły mechanizmy, które w pewnych zakresach zakłóceń przekazu pozwalają jednak odbiorcy komunikatu na zrozumienie jego treści, zgodnie z intencją nadawcy. Mechanizmy te polegają na wpleceniu dodatkowych informacji na różnych poziomach systemu komunikacyjnego i nazywane są redundancją. Przykłady takich redundancji dla różnych warstw systemów porozumiewania to: zwiększenie głośności przekazu czy jego spowolnienie w przypadku komunikacji głosowej, budowa języka np. w sposób umożliwiający odczytanie przekazu nawet po usunięciu z niego niektórych samogłosek w przypadku komunikacji pisanej, czy też wreszcie możliwość dopowiedzenia sobie pewnych zniekształconych fragmentów komunikatu na podstawie wcześniejszego kontekstu. Więcej, zarówno o pozytywnych, jak i negatywnych skutkach obecności redundancji w językach naturalnych, znaleźć można w rozważaniach autorki na temat rozwoju języków graficznych (Kamińska, 2017a), natomiast wielowymiarowy opis tego zjawiska przedstawiają Ernst Wit oraz Marie Gillette w swoim raporcie *What is Linguistic Redundancy?* (Wit & Gillette, 1999).

Warto tutaj zauważyć, że jeśli w przekazie tekstowym tworzonym odręcznie pewne znaki mogą zostać słabiej zapisane lub mieć zdeformowane kształty, to przekaz taki (oczywiście do pewnego stopnia jego deformacji) może w dalszym ciągu zostać poprawnie odczytany, choć być może wydłuży to czas jego odczytu. Sytuacja zmienia się w przypadku interakcji człowieka z systemami komputerowymi (gdzie na poziomie tabel kodowania znaków tekstowych (ang. *character sets*) unika się redundancji), a konkretnie zapisywania/wprowadzania informacji za pomocą klawiatur komputerowych. W takich przypadkach naciśnięcie klawisza „tylko trochę" lub „trochę obok" odniesie skutek zapisania bądź całkowitego pominięcia danej litery lub zapisania danej bądź sąsiedniej litery. O ile nawet w przypadku

takich błędów przekłamana informacja często może być jeszcze poprawnie zinterpretowana przez człowieka, to klasyczne metody przetwarzania informacji w systemach komputerowych, opierające się na ostrych kryteriach podobieństw czy przynależności elementów do zbiorów, mogą uniemożliwić poprawne funkcjonowanie systemów w takich właśnie sytuacjach.

Niniejsze rozważania dotyczą bibliograficznych baz danych, zwłaszcza tych o „tradycyjnej organizacji przestrzeni informacyjnej". Zagrożenia i problemy występujące w takich systemach zostały przedstawione na przykładach wybranych krajowych baz bibliograficznych (BazTech, CYTBIN) w osobnym opracowaniu (Kamińska, 2017d), gdzie wykazuje, że problemy te wynikają z nakładania się przyjętych założeń co do modelu przestrzeni informacyjnej oraz błędów literowych i niekonsekwencji w stosowaniu formatów zapisów, podczas opisywania jednostek bibliograficznych, które mogą prowadzić do przechowywania różnych informacji opisujących te same jednostki. Przypadki, wpisujące się w wyżej opisaną sytuację, mogą prowadzić do zaburzeń w realizacji funkcji informacyjnych systemów bibliograficznych oraz do problemów z analizami bibliometrycznymi realizowanymi w oparciu o dane zgromadzone w takich systemach. Przykłady tych ostatnich przedstawia autorka w osobnych opracowaniach dokumentujących uzyskane wyniki badań bibliometrycznych (Kamińska, 2017c; Kamińska, 2017e) przeprowadzonych na podstawie danych zgromadzonych w krajowej bibliograficznej bazie danych CYTBIN[1].

W dalszej części rozważań przedstawione zostaną na przykładach propozycje wykorzystujące miary podobieństw łańcuchów znakowych do wykrywania duplikatów rekordów bibliograficznych. Problem zjawiska powielania rekordów został dostrzeżony już wraz z rozwojem praktycznych zastosowań systemów zarządzania bazami danych, zaś sposoby identyfikacji powielonych rekordów i przeciwdziałania temu zjawisku badane są już od kilkunastu lat, głównie przez badaczy zagranicznych. Szeroki przegląd stosowanych praktyk i trendów prezentują chociażby Gu, Baxter, Vickers i Rainsford (Gu et al., 2003), podczas gdy Jiang, Lin, Meng, Yu, Cohen i Smalheiser przedstawiają koncepcję deduplikacji rekordów bibliograficznych pochodzących z wielu baz źródłowych, a realizowaną w trybie natychmiastowym (ang. *online*) na żądanie użytkownika (Jiang et al., 2014). Dla zwiększenia wydajności przetwarzania, autorzy przyjęli tam możliwość ograniczenia przestrzeni poszukiwań do podzbiorów/partycji opartych na roku wydania poszczególnych jednostek bibliograficznych. Temat deduplikacji rekordów bibliograficznych składowanych w formacie UNIMARC poruszali również Nuno Freire, José Borbinha i Pável Calado (Freire et al., 2007).

[1] Baza danych dostępna jest pod adresem http://www1.bg.us.edu.pl/bazy/cytbin/, natomiast pod adresem http://www1.bg.us.edu.pl/bazy/cytbin/opis_cytbin.html znaleźć można jej opis, krótką charakterystykę oraz listę współtworzących ją osób.

Niestety, w kraju, w dziedzinie bibliograficznych baz danych, zagadnienia te nie wydają się powszechnie znane, co potwierdzają liczne przykłady anomalii w postaci powielonych danych występujących nie tylko pomiędzy, ale co gorsza, w obrębie poszczególnych systemów. Małe zainteresowanie tą klasą zagadnień w krajowym czasopiśmiennictwie dotyczącym bibliograficznych baz danych (stąd rozdział kolejny stanowi krótkie wprowadzenie w pryncypia miar podobieństw łańcuchów znakowych) oraz doświadczenia nabyte podczas budowy własnej bazy bibliograficznej, których opis znaleźć można w osobnym opracowaniu (Kamińska, 2017b), były motywacjami do podjęcia badań przedstawionych w ramach niniejszych rozważań.

W artykule dotyczącym wyszukiwania powielonych opisów bibliograficznych Adrian Drabik proponuje natomiast metodę bazującą na porównywaniu częstości wystąpień znaków do wykrywania podobnych opisów bibliograficznych. Motywacją autora było obniżenie złożoności obliczeniowej, jednak jak on sam zauważa „proponowane algorytmy są w pierwszej kolejności nieprecyzyjne, stąd ich zastosowanie jest uzasadnione jedynie w szczególnych przypadkach" (Drabik, 2016, s. 78). Warto również zauważyć, że o ile jeszcze w przypadku słów składających się z liter zbudowany wektor częstości mógłby w dość selektywny sposób opisywać dane słowo (choć opisana metoda nie bazuje na słowach tylko całych fragmentach opisów tekstowych), to w przypadku liczb (a mamy z nimi do czynienia w przypadku danych bibliograficznych, np. w numerach stron czy roku wydania) jest już dużo gorzej. Dzieje się tak dlatego, że liczby budowane są z bardziej ograniczonego podzbioru znaków (czyli cyfr) oraz, co ważniejsze, posiadają więcej poprawnych form sprowadzających się do tego samego wektora częstości. Na przykład liczba 13313 „częstościowo" będzie identyczna zarówno z liczbą 13133 jak i z liczbą 33311 oraz wieloma innymi. Wydaje się jednak, że prawdopodobieństwo wprowadzenia liczby 13133 zamiast 13313 na skutek powstania „błędu typograficznego" jest dużo wyższe (a więc liczby te powinny być „bardziej do siebie podobne") niż liczby 33311, czego metoda częstościowa nie jest w stanie uwzględnić. Dlatego też autorka proponuje bardziej podstawowe podejście, bazujące na miarach podobieństw łańcuchów znakowych, które uwzględniają relacje porządku znaków w tych łańcuchach, co pozwala na skuteczną identyfikację rekordów podobnych, których różnice wynikają z błędów typograficznych. Metoda została zweryfikowana w praktyce oraz zilustrowana na przykładzie danych zaczerpniętych z bibliograficznej bazy CYTBIN.

PODSTAWOWE MIARY PODOBIEŃSTW ŁAŃCUCHÓW ZNAKOWYCH

Jak już wspomniano, klasyczne metody przetwarzania informacji w systemach komputerowych opierają się na logice dwuwartościowej, której wykorzystanie implikuje najczęściej ostre granice podobieństwa dwóch ele-

mentów (albo są one identyczne, albo różne), czy przynależności do zbiorów (element albo należy do zbioru, albo do niego nie należy). Człowiek na co dzień posługuje się pojęciami wykraczającymi poza tak sztywno przyjęte ramy – przykładowo: „Dzisiaj trochę pada" lub „Ten pies jest prawie identyczny jak pies sąsiada". Potrzeba przeniesienia modeli pojęciowych do świata komputerów zaowocowała rozwojem koncepcji, jak zbiory rozmyte. W ich przypadku zagadnienia przynależności danego elementu do zbioru nie sprowadza się do dwóch możliwych odpowiedzi („tak" lub „nie"), ale określa się funkcję przynależności do zbioru przyjmującą wartości rzeczywiste z przedziału <0,1>. Podążając za przywołanym przykładem można się więc tutaj posłużyć sformułowaniem: „Ten pies jest podobny do psa sąsiada w stopniu 0,9 (90%)".

Szczególnym przypadkiem powyższych rozważań jest określanie stopnia podobieństwa między dwoma łańcuchami znakowymi. Poniżej przedstawione zostaną podstawowe, wybrane metody skonstruowane specjalnie w tym celu. Zostaną one opisane z perspektywy użytkowej, zostawiając perspektywę implementacyjną oraz formalne ich opisy specjalistom od algorytmiki. Warto zauważyć, że używanym w tej dziedzinie terminem odwrotnym do podobieństwa jest odległość – ciągi znaków są tym bardziej od siebie odległe, im mniejsze jest podobieństwo między nimi.

Odległość Hamminga (ang. *Hamming distance*) nazwę swoją wywodzi od jej twórcy, który pierwszy raz opisuje ją w swoim artykule z 1950 r. (Hamming, 1950). Mierzona być może dla ciągu znaków o równej długości. Wyraża się ona liczbą pozycji, na jakich symbole dwóch ciągów znaków są różne. Mówiąc innymi słowy, mierzy ona minimalną liczbę podmian symboli tak, aby przekształcić jeden ciąg w drugi lub, jeszcze inaczej, minimalną liczbę błędów transmisji, które musiałyby zaistnieć, aby zniekształcić jeden ciąg do postaci drugiego. Biorąc pod uwagę wymóg równych długości mierzonych ciągów, metoda ta może znaleźć ograniczone zastosowanie w deduplikacji rekordów bibliograficznych (choć i w tej dziedzinie występują takie atrybuty opisowe – np. odpowiednio dziesięcio- oraz trzynastocyfrowe numery ISBN), ale została przywołana ze względu na jej obrazową naturalność i intuicyjność. Wartości tej miary dla kilku wybranych ciągów znaków przedstawiono w tabeli 1.

Tabela 1

Przykładowe wartości odległości Hamminga

Łańcuch A	Łańcuch B	Wartość miary
Paweł	Gaweł	1
234543	234567	2
0011101	0001001	2

Jedną z najbardziej znanych miar odległości między łańcuchami znaków jest również dość elementarna koncepcyjnie odległość Levenshteina (ang. *Levenshtein distance*) opisana przez jej autora w artykule z 1965 r. (Левенштейн, 1965), a stanowiąca uogólnienie przedstawionej wcześniej odległości Hamminga. Wartość tej miary wyliczona dla dwóch ciągów wskazuje minimalną liczbę operacji edycyjnych (zastąpienia, dopisania bądź usunięcia znaku), niezbędną do przekształcenia jednego ciągu znaków w drugi. Widać tutaj, że skoro zbiór operacji, poza zastępowaniem, rozszerzony został o operację dodania i usunięcia znaku, to porównywane łańcuchy znaków nie muszą już być tej samej długości. Wartości tej miary dla kilku wybranych ciągów znaków przedstawiono w tabeli 2.

Tabela 2

Przykładowe wartości odległości Levenshteina

Łańcuch A	Łańcuch B	Wartość miary
Paweł	Gaweł	1
234543	234567	2
0011101	0001001	2
Patrycja	Alicja	5
134567	1234567	1
1234567	1324567	2

Przedostatni wiersz tabeli z powyższego przykładu uwidacznia zalety tej miary – mimo że zgodność dwóch łańcuchów znakowych występuje tylko na pierwszej pozycji, to wartość użytej miary i tak wskazuje na bardzo duże ich podobieństwo. Widzimy, że pierwszy łańcuch różni się od drugiego pominięciem jednego znaku lub, drugi od pierwszego, dodaniem jednego znaku. Są to bardzo częste przypadki błędów typograficznych występujących podczas wprowadzania danych za pomocą klawiatury komputerowej i dlatego miara ta okazuje się bardzo skuteczna dla wykrywania rozbieżności rekordów tym właśnie spowodowanych.

Ostatnią z przedstawianych, w ramach niniejszych rozważań, miar odległości między łańcuchami znakowymi jest odległość Jaro-Winkler (ang. *Jaro-Winkler distance*). Stanowi ona udoskonaloną (o preferowanie znaków wspólnych na początkach ciągów) metodę (opartą o odległość Jaro) nie bazującą na tradycyjnych „odległościach edycyjnych". Metoda ta została opracowana specjalnie dla celów deduplikacji rekordów w bazach danych z myślą o jej wykorzystaniu, zwłaszcza w przypadku krótkich łańcuchów znakowych. Artykuł ją przedstawiający opublikowany został w 1989 r. (Jaro, 1989). Wartości tej miary dla kilku wybranych ciągów znaków przedstawiono w tabeli 3. Warto zwrócić uwagę, że w odróżnieniu od poprzed-

Tabela 3

Przykładowe wartości odległości Jaro-Winkler

Łańcuch A	Łańcuch B	Wartość miary
Paweł	Gaweł	0,86(6)
234543	234567	0,86(6)
0011101	0001001	0,7943
Patrycja	Alicja	0,625
Patrycja	Patrycji	0,95
Patrycja	patrycja	0,91(6)
134567	1234567	0,9571
1234567	1324567	0,9571

nich, wartości tej miary reprezentują stopień podobieństwa (a nie różnice) i normalizowane są do przedziału <0,1>.

Na przykładzie par ciągów („Patrycja", „Patrycji") oraz („Patrycja", „patrycja") zaobserwować można wyraźną preferencję zgodności początkowych ciągów znaków. Przykład ten pokazuje również, że wykorzystywana implementacja metody rozróżnia wielkości znaków.

Przedstawione miary podobieństw (lub różnic) łańcuchów znakowych to tylko nieliczne spośród najbardziej popularnych i najczęściej używanych miar w zastosowaniach deduplikacji danych. Bardziej obszerne rozważania na temat tych i innych metod wraz z omówieniem ich efektywności prowadzą Cohen, Ravikumar i Fienberg (Cohen et al., 2003). Jako że złożoność obliczeniowa niektórych miar jest znaczna, to przy dużych zbiorach danych bibliograficznych wybór konkretnej z nich może być również dyktowany względami ograniczeń czasowych na realizację procesu deduplikacji. Dlatego też rozważania na temat opisywanej metody zakończone zostaną dygresją na temat wydajności obliczeniowej oraz jej skalowalności w odniesieniu do przyrostu bazy danych, spowodowanego napływem coraz to nowych danych bibliograficznych.

STUDIUM WYSTĘPUJĄCYCH NIEPRAWIDŁOWOŚCI

W zależności od architektury informacyjnej przyjętej dla konkretnej bibliograficznej bazy danych, istnieć może kilka potencjalnych źródeł powstawania redundantnych rekordów. Materiał badawczy wykorzystany w ramach tej pracy zaczerpnięty został z jednej z krajowych bibliograficznych baz danych – CYTBIN. Analizowane dane pozyskano wprost z publicznie dostępnej aplikacji WWW. Baza danych posiada tradycyjną architekturę informacyjną, co oznacza, że dane o bibliograficznych jednostkach cytujących i cytowanych przechowywane są w oddzielnych składnicach, z dodatkowym podziałem

na typy jednostek cytowanych. Szersze omówienie przyjętego tam modelu danych oraz przyczyn zjawiska powstawania duplikatów rekordów znaleźć można w opracowaniu na temat problemów dotykających tradycyjne bibliograficzne bazy danych i metodzie pozwalającej na ich unikanie (Kamińska, 2017d). Następnie pozyskanymi danymi zasilono relacyjną bazę danych Oracle, a do obliczania miar podobieństwa łańcuchów znakowych wykorzystano wbudowany w tę bazę danych pakiet programowy UTL_MATCH zawierający implementację wielu ze wspomnianych wcześniej miar. Można przypuszczać, że implementacje algorytmów miar dostarczane przez poszczególnych dostawców mogą się nieznacznie różnić. Dotyczy to w szczególności miary Jaro-Winkler, dla której można sobie wyobrazić implementacje uwzględniające w różnym stopniu wpływ pozycji różnych znaków na jej wartość.

Niniejsze rozważania ogranicza się do jednostek bibliograficznych (w tym tytułu i roku wydania) oraz ich autorów (w tym nazwiska i imion), gdyż model metadanych przyjęty dla analizowanej bazy danych jest stosunkowo prosty, natomiast w bardziej rozbudowanych/ustrukturalizowanych systemach mogą to być również encje instytucji sprawczych, wydawnictw, jednostek naukowych i inne.

W przypadku danych o autorach nie sposób nie zauważyć, że struktura rekordów ich opisujących nie jest złożona. Nawet w przypadku ustrukturalizowanych informacji mamy tutaj do czynienia najczęściej z trzema atrybutami: nazwisko, imię oraz drugie imię. Chociaż z wymienionych atrybutów nazwisko wydaje się najbardziej definiować konkretną osobę, to z oczywistych względów jest ono do tego niewystarczające. Dlatego też, chcąc stosować prostą jednoetapową metodę wykrywania powielonych rekordów, najprościej jest połączyć wszystkie trzy atrybuty w jeden łańcuch znakowy, rozpoczynający się nazwiskiem, a kończący drugim imieniem i użyć miary Jaro-Winkler jako faworyzującej podobieństwo znaków początkowych. Poza przypadkami błędów literowych, powinno to również pomóc w znalezieniu rekordów opisujących te same osoby raz przy użyciu nazwiska i pełnego imienia, a innym razem jedynie nazwiska i inicjału. Warto zwrócić uwagę na badania nad ulepszaniem miary Jaro-Winkler (Dressler & Ngonga Ngomo, 2017). Kiedy dysponujemy ich implementacją – dobór parametrów, np. do specyfiki konkretnych metadanych stanowi jeden z elementów strojenia całej metody.

Kilka pierwszych (względem miary Jaro-Winkler) przykładowych rekordów znalezionych za pomocą powyżej opisanej metody przedstawiono w tabeli 4.

Widzimy tutaj dwie zasadnicze kategorie różnic pomiędzy zidentyfikowanymi rekordami podobnymi: błędy typograficzne w nazwiskach oraz niekonsekwentne używanie znaku kropki po inicjale imienia. Drugi przypadek może zostać zidentyfikowany (i skorygowany automatycznie bez potrzeby ingerencji operatora) bez uciekania się do bardziej kosztownych obliczeniowo miar podobieństw, np. przez użycie zwykłego porównania

Tabela 4

Najbardziej podobne rekordy autorów po zastosowaniu miary Jaro-Winkler

Autor A	Autor B	Wartość Jaro-Winkler
Pasztaleniec-Jarzyńska J.	Pasztelaniec-Jarzyńska J.	0,9923076923076923
Pasztaleniec-Jarzyńska J	Pasztaleniec-Jarzyńska J.	0,9923076923076923
Nikodemska-Wołowik A.M.	Nikodemska-Wołownik A.M.	0,992
Majkowska-Aleksiewicz A	Majkowska-Aleksiewicz A.	0,9916666666666667
Siniarska-Czaplicka J	Siniarska-Czaplicka J.	0,990909090909091
Korczyńska-Derkacz M	Korczyńska-Derkacz M.	0,990909090909091
Korczyńska-Dekacz M.	Korczyńska-Derkacz M.	0,990909090909091
Rusińska-Giertych H	Rusińska-Giertych H.	0,9904761904761905
Bednarek-Michalska B	Bednarek-Michalska B.	0,9904761904761905
Kamińska-Czubała B	Kamińska-Czubała B.	0,9904761904761905
Morsztynkiewiczowa I	Morsztynkiewiczowa I.	0,9904761904761905
Woźniak-Kasperek J	Woźniak-Kasperek J.	0,99
Żbikowska-Migoń A	Żbikowska-Migoń A.	0,99
Busse-Turczyńska E	Busse-Turczyńska E.	0,99
Kurek-Kokocińska S	Kurek-Kokocińska S.	0,99
Okoń-Horodyńska E	Okoń-Horodyńska E.	0,99

łańcuchów z użyciem operacji usunięcia każdego znaku kropki występującego w rekordzie. Analizy tą metodą wykazały, że przypadek ten dotyczył 3,9% z wszystkich rekordów zawierających dane dotyczące autorstwa publikacji. Również częstym przypadkiem, możliwym do wykrycia bez uciekania się do korzystania z miar podobieństw, jest niekonsekwentny zapis imion dla danego autora – raz jako imienia w postaci pełnej, a innym razem jedynie jako inicjału. Innym przykładem możliwym do wykrycia prostymi sposobami jest niekonsekwentne używanie wielkich liter.

Wykorzystywanie miar podobieństw do w pełni automatycznej deduplikacji rekordów jest obarczone ryzykiem błędnej deduplikacji rekordów podobnych (tzw. efekt „false positives"), opisujących jednak różne byty i dlatego powinno się je wykorzystywać raczej tylko do wskazania operatorowi rekordów kandydujących, zaś ten biorąc na siebie rolę arbitra, powinien podejmować każdorazowe decyzje.

Na rysunku 1 przedstawiono histogram miary Jaro-Winkler dla analizowanych rekordów. Aby nie popełnić błędów w interpretacji wyników, należy zwrócić uwagę na logarytmiczną skalę osi rzędnych.

Analizowana baza autorów liczy w przybliżeniu 9 tys. rekordów. Wynika z tego, że liczba par wszystkich rekordów wynosi (z pominięciem par „przemiennych") 40 mln (n^2/2-n/2), co stanowi znaczny wolumen

Rys. 1. Histogram miary Jaro-Winkler dla analizowanych rekordów danych

danych, zwłaszcza w kontekście manualnej ich weryfikacji. Jednak patrząc na rysunek 1 widzimy, że dla wartości podobieństw większych od 90% liczba rekordów oscyluje wokół rzędu wielkości 1 tys. W analizowanej bazie całkowita liczba rekordów z miarą podobieństwa, większą lub równą 90%, wyniosła 5406, co stanowi 0,13‰ liczby wszystkich możliwych par i 60% liczby wszystkich rekordów opisujących autorów. Nawet

Tabela 5

Podobne rekordy autorów względem miary Jaro-Winkler

Autor A	Autor B	Wartość Jaro-Winkler
Abraham F. D.	Abrams D.	0,9008547008547009
Adamczewski Z.	Adamiec M.	0,9085714285714285
Adamczewski Z.	Adamiec W.	0,9085714285714285
Adamczewski Z.	Adamski F.	0,9085714285714285
Adamczewski Z.	Adamski S.	0,9085714285714285
Aleksandrow A. D.	Aleksandrowicz R.	0,9058823529411764
Aleksandrow A. D.	Aleksandrowicz T.	0,9058823529411764
Aleksandrow A. D.	Aleksandrowska E.	0,9058823529411764
Aleksandrowicz R.	Aleksandrowska E.	0,9058823529411764
Aleksandrowicz T.	Aleksandrowska E.	0,9058823529411764
Allemang D.	Allen B.	0,9022727272727272
Allen B.	Allen M. R.	0,9022727272727272
Andersen T.	Anderson S.J.	0,9020979020979021
Anderson J. D.	Andersson K.	0,9095238095238094
Anderson R.	Andrews R.	0,9083164983164983
Anderson T. H.	Andersson K.	0,9095238095238094

dla tak przyjętego progu, przykładowe wartości w tabeli 5 pokazują, że większość rekordów jest jeszcze od siebie „bardzo odległa" i w praktyce próg też można by znacznie podwyższyć, czyniąc zbiór „par podejrzanych" o wiele mniejszym. Trzeba jednak zauważyć, że zbyt mocne podnoszenie progu podobieństwa może skutkować wykluczeniem ze zbioru „par podejrzanych" tych, które reprezentują jednak osobne byty (tzw. efekt „false negatives").

Należy tutaj wyraźnie zaznaczyć, że każdy system bibliograficznej bazy danych jest inny wraz ze swoimi specyficznymi zaletami i specyficznymi wadami. W analizowanych danych poza wykrytymi błędami typograficznymi bardzo często występował problem niekonsekwentnego używania znaków kropki po inicjałach. W innych systemach może być to stosowanie wielokrotnych spacji, średników czy innych znaków, dlatego trudno tutaj mówić o uniwersalnym algorytmie deduplikacji danych. Ponieważ klasy występujących anomalii w danych i ich charakter ilościowy zależą również od wad projektowych konkretnego systemu, zatem i opracowywanie mechanizmów deduplikacji powinno być wykonywane indywidualnie, po wcześniejszym zapoznaniu się z typami błędów kumulujących się w danej bazie danych. Można natomiast pokusić się o przedstawienie ogólnego podejścia do projektowania takich mechanizmów. Składa się ono z następujących kroków:

1. Zapoznanie się z wolumenem danych i typologią występujących anomalii.
2. Zidentyfikowanie typów anomalii, których usunięcie może być obsłużone bez udziału operatora i zaprojektowanie „ostrych" (ang. crisp) algorytmów ich automatycznego usuwania.
3. Ustalenie progu wartości dla użytych miar podobieństwa, na podstawie których oznaczane będą kandydujące do deduplikacji rekordy (początkowo, podczas wdrażania mechanizmów deduplikacji, warto te progi ustawić wysoko, co pozwoli skupić się na najbardziej „oczywistych przypadkach", zaś później rozważyć ich stopniowe obniżanie).
4. Uwzględniając wolumen danych, dostępność mocy obliczeniowych oraz możliwość zaangażowania operatorów, zaplanować cykliczny harmonogram oznaczania rekordów kandydujących do deduplikacji lub zaimplementować mechanizmy weryfikacji podobieństwa rekordu (do rekordów już zgromadzonych) w chwili jego rejestrowania.

Prowadząc rozważania na temat deduplikacji rekordów opisujących autorów, nie sposób nie zauważyć, że istnieją różni autorzy o tych samych imionach i nazwiskach. Dobrze zaprojektowana bibliograficzna baza danych powinna umożliwiać poprawną obsługę takich przypadków i pozwalać na rozróżnienie poszczególnych autorów za pomocą nadanych im identyfikatorów. Rekordy takie w oczywisty sposób stanowić będą potencjalnych kandydatów do deduplikacji, jednak system po pierwszej decyzji

operatora o ich niescaleniu, powinien tę decyzję zapamiętać i nie oznaczać już tych rekordów w przyszłości.

Z odwrotną sytuacją mamy do czynienia w przypadku zmiany nazwiska bądź imienia autora. Deduplikacja takich rekordów może prowadzić do zafałszowania bibliografii załącznikowej, natomiast brak takiej deduplikacji prowadzi do zafałszowania wyników analiz bibliometrycznych czy zaburzenia mechanizmów wyszukiwawczych. Dobrze zaprojektowany system powinien uwzględniać możliwość zmiany danych autora w czasie, a jednocześnie pozwalać na jednoznaczne rozpoznanie takiej osoby, bez względu na różne dane ją opisujące, poprzez nadany wcześniej identyfikator.

Kolejnym rodzajem danych, w których powszechnie wykrywa się anomalie, są rekordy „bardziej złożone strukturalnie" (tj. takie, w których atrybuty są różnych typów danych, jak na przykład daty, liczby i łańcuchy znakowe) – np. opisywanych jednostek bibliograficznych. Podejście połączenia ich zakresu informacyjnego do jednego łańcucha znakowego

Tabela 6

Podobne rekordy tytułów względem miary Jaro-Winkler

Tytuł A	Tytuł B	Wartość Jaro-Winkler
Bibliografia	Bibliografia	1,0
Bibliografia polska. T. 15	Bibliografia polska. T. 5	0,9923076923076923
Bibliography of Otlet's works and secondary sources	Bibliography of Otlet`s works and secondary sources	0,9921568627450981
Bibliografia zawartości Przeglądu Bibliotecznego 1977-1996 (R. 45-64)	Bibliografia zawartości „Przeglądu Bibliotecznego" 1977-1996 (R.45-64)	0,9916275430359938
Bibliografia Wydawnictw Ciągłych 1982	Bibliografia Wydawnictw Ciągłych 1983	0,9897435897435898
Bibliografia Wydawnictw Ciągłych 1983	Bibliografia Wydawnictw Ciągłych 1984	0,9897435897435898
Bibliografia Wydawnictw Ciągłych 1981	Bibliografia Wydawnictw Ciągłych 1984	0,9897435897435898
Bibliografia Wydawnictw Ciągłych 1981	Bibliografia Wydawnictw Ciągłych 1983	0,9897435897435898
Bibliografia Wydawnictw Ciągłych 1981	Bibliografia Wydawnictw Ciągłych 1982	0,9897435897435898
Bibliografia Wydawnictw Ciągłych 1982	Bibliografia Wydawnictw Ciągłych 1984	0,9897435897435898
Bibliograficznych ksiąg dwoje. T. 2	Bibliograficznych ksiąg dwoje. T. 1	0,9888888888888889
Bibliografia polska. T. 27	Bibliografia polska. T. 22	0,9846153846153847
Bibliografia polska. T. 17	Bibliografia polska. T. 16	0,9846153846153847
Bibliografia polska. T. 15	Bibliografia polska. T. 16	0,9846153846153847

nie odniosłoby oczekiwanego skutku, zwłaszcza z wykorzystaniem miary Jaro-Winkler, gdyż trudno byłoby uporządkować atrybuty składające się na cały opis jednostki w kolejności ich ważności, a jak zauważono wcześniej, zgodność szczególnie początkowych liter ma dla tej miary bardziej istotne znaczenie. Choć naturalnym rozwiązaniem mogłoby wydawać się przyjęcie tytułu opisywanej jednostki za pierwszy atrybut, to podejście takie skutkować by mogło parowaniem rekordów o tych samych tytułach. W przypadku cyklu publikacji, dla których tytuł kończony jest numerem części, będzie to prowadziło do wielu błędnych wyników. Obrazuje przykład przedstawiony w tabeli 6.

Mimo wysokich miar podobieństwa widać, że tylko pary rekordów z wiersza 3 i 4 są potencjalnymi kandydatami do deduplikacji (ze względu na występowanie błędów – odpowiednio są to inne znaki apostrofu i brak dodatkowego znaku spacji). Tytuły pozostałych rekordów opisują różne jednostki bibliograficzne (jak się okazało po weryfikacji, również pozycje z pierwszego wiersza różnią się datą wydania). Widać, że aby oznaczanie rekordów kandydujących do deduplikacji metodami określania odległości pomiędzy łańcuchami znakowymi było użyteczne, potrzebna jest inna, bardziej złożona miara. Poza tytułem, dobrym wyróżnikiem jednostki bibliograficznej wydaje się rok wydania i na jego przykładzie zostanie zaprezentowana przedstawiana koncepcja (w innych przypadkach w skład miar złożonych wchodzić mogą również numery stron, instytucje wydawnicze i inne).

Wybierając znormalizowane implementacje miar podobieństw (czyli takie, które dla łańcuchów najbardziej odległych wynoszą 0, zaś dla łańcuchów identycznych 1), można zbudować miarę złożoną, gdzie dla każdego wybranego atrybutu użyje się danej miary, zaś końcowy wynik stanowił będzie sumę wartości poszczególnych miar, z uwzględnieniem ustanowionych odpowiednio wag. Dla omawianego przykładu wybrano miarę Jaro-Winkler dla tytułów oraz znormalizowaną miarę odległości edycji (ang. *edit distance*) dla lat publikacji, z równymi wagami (przyjęto wartości 1/2), otrzymując wyniki przedstawione w tabeli 7. Natomiast ogólna postać miary złożonej wyraża się poniższym wzorem:

$$M = \sum_{i=1}^{n} w_i \cdot m_i$$

gdzie m_i jest wartością miary wybranego typu dla i-tego atrybutu, zaś w_i wagą ustaloną dla i-tego atrybutu.

Z danych zebranych w tabeli 7 wynika wyraźnie, że dzięki zastosowaniu miary złożonej wyeliminowano problematyczne rekordy występujące w tabeli 6, a w ich miejsce pojawiły się nowe rekordy, niektóre również źle dopasowane, ale w mniej oczywisty sposób. Jak zaznaczono wcześniej,

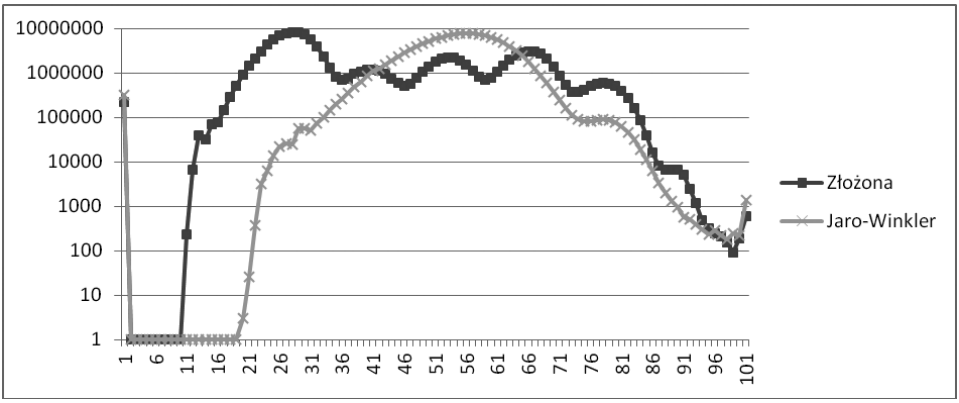
Tabela 7

Podobne rekordy tytułów po zastosowaniu miary złożonej

Tytuł A	Rok	Tytuł B	Rok	Miara
Bibliografia niemieckich bibliografii dotyczących Polski 1900-1958	1960	Bibliografia niemieckich bibliografii dotyczących Polski 1900-1958	1960	200
Bibliografia zawartości Przeglądu Bibliotecznego 1977-1996 (R. 45-64)	1999	Bibliografia zawartości „Przeglądu Bibliotecznego" 1977-1996 (R.45-64)	1999	199
Bibliography of Otlet's works and secondary sources		Bibliography of Otlet`s works and secondary sources		199
Bibliografia Wydawnictw Ciągłych 1981	1984	Bibliografia Wydawnictw Ciągłych 1982	1984	198
Bibliografia regionalna w warunkach automatyzacji	1995	Bibliografie regionalne w warunkach automatyzacji	1995	192
Bibliografia Bibliografii i Nauki o Książce		Bibliografia Bibliografii Polskich		191
Bibliografia publikacji pracowników WSP w Kielcach za lata 1977-1978 (z uzupełnieniami do roku 1976)	1984	Bibliografia publikacji pracowników WSP w Kielcach 1979-1982	1984	191
Bibliografia Bibliografii Polskich		Bibliografia Bibliografii i Nauki o Książce		190
Bibliographie		Bibliologia		190
Bibliografia publikacji pracowników Uniwersytetu Warszawskiego	2005	Bibliografie publikacji pracowników instytucji naukowych w Polsce	2005	189
Bibliografia Regionalna. Informacja o pracach Zespołu ds. Bibliografii Regionalnej ZG SBP	1999	Bibliografie regionalne : informacja o pracach Zespołu ds.. Bibliografii Regionalnej ZG SBP	1999	189
Bibliografia bibliografii.	1986	Bibliografia	1986	189
Bibliographie		Biblioteka		188
Bibliography (Soshigaku)	1983	Bibliografia	1983	187

deduplikacja realizowana przy użyciu miar nieostrych powinna się odbywać przy udziale czynnika ludzkiego, zaś konstruowanie takich a nie innych miar służy do tego, aby ten czynnik ludzki jak najefektywniej wykorzystać.

Na rysunku 2 przedstawiono histogramy dla analizowanych artykułów wykreślone dla zaproponowanej miary Jaro-Winkler wyliczanej tylko w odniesieniu do tytułów oraz miary złożonej. Liczba analizowanych artykułów bliska była 16,5 tys., zatem liczba potencjalnych par wynosi prawie 136 mln. Można zauważyć, że przebieg histogramu dla miary złożonej wskazuje na większą liczbę par słabiej ocenionych, zaś w części środkowej



Rys. 2. Histogram miary Jaro-Winkler i miary złożonej dla analizowanych rekordów artykułów

jego przebieg jest bardziej wyrównany i oscyluje wokół stałej wartości. Na podstawie jego kształtu można by rozpocząć analizę podobieństw na poziomie wartości miary 80, a następnie w zależności od obserwowanej częstości błędnie sugerowanych deduplikacji przesuwać się w prawą lub lewą stronę osi, ustalając ostatecznie roboczy próg wartości.

Jak już wspomniano wcześniej, w bibliograficznych bazach danych mogą być również wyodrębnione informacje o instytucjach sprawczych, wydawnictwach, jednostkach naukowych i innych, które również mogą być podatne na zjawisko duplikacji rekordów. W zależności od rozmiaru informacyjnego konkretnych encji deduplikacja realizowana w ich ramach może być przeprowadzana zgodnie z koncepcją, przedstawioną dla deduplikacji autorów lub też dla deduplikacji jednostek bibliograficznych.

Prowadząc rozważania dotyczące deduplikacji rekordów, czy szerzej dbałości o wiarygodność i użyteczność gromadzonych i przetwarzanych danych bibliograficznych, nie sposób nie wspomnieć o *stricte* bibliometrycznych miarach podobieństw między jednostkami blibliograficznymi, jak np. metoda powiązań bibliograficznych, opracowana przez Michaela Kesslera w 1963 r. (Kessler, 1963), a której szerszego omówienia w języku polskim dokonała Irena Marszakowa-Szajkiewicz (Marszakowa-Szajkiewicz, 2009, s. 136-137). W trakcie realizacji własnych badań bibliometrycznych na danych udostępnionych przez jedną z krajowych baz bibliograficznych (Kamińska, 2017e), okazało się, że korzystając ze wspomnianej metody, autorka niejako przy okazji zidentyfikowała zduplikowane rekordy opisujące dane jednostki bibliograficzne (Kamińska, 2017c). Warto jednak zauważyć, że metoda ta (jako porównująca zbiory wspólne bibliografii załącznikowych) jest tym skuteczniejsza, im więcej pozycji bibliograficznych zawierają dane jednostki oraz że możliwość wykorzystania tej metody warunkowana jest gromadzeniem pełnych opisów bibliografii załącznikowych dla porównywanych jednostek.

ZŁOŻONOŚĆ OBLICZENIOWA

Proponowana metoda została zweryfikowana w praktyce nie tylko pod kątem spełniania potrzeb funkcjonalnych, ale również kryteriów wydajnościowych. Dla bazy danych zawierającej opisy blisko 16,5 tys. jednostek bibliograficznych posadowionej na biurowym systemie komputerowym przeciętnej mocy obliczeniowej (Intel Core i5), wyliczenie i uporządkowanie (sortowanie malejące) podobieństwa wszystkich rekordów bibliograficznych do zadanego, metodą złożoną opartą na kilku atrybutach trwało poniżej 0,7 s (należy się liczyć z tym, że czas ten zawiera również pewien stały narzut, np. na analizę składniową czy komunikację pomiędzy komputerem klienckim i serwerem). Wydaje się, że jest to wartość akceptowalna dla wdrożenia metody w trybie proaktywnym, czyli umożliwiającym podpowiadanie rekordów podobnych, jeszcze przed zatwierdzeniem właśnie wprowadzanego przez operatora opisu bibliograficznego, celem podjęcia decyzji o ewentualnym scaleniu zapisów. Porównanie podobieństwa pomiędzy wszystkimi zgromadzonymi opisami bibliograficznymi w bazie danych (czyli każdy z każdym) zajmowało natomiast średnio 56 min.

Warto przyjrzeć się złożoności obliczeniowej (czasowej) dwóch powyższych operacji, czyli ich zapotrzebowaniu na moc obliczeniową w zależności od skali problemu (liczby rekordów bibliograficznej bazy danych). Złożoność obliczeniową określa się zawsze w kontekście operacji dominującej realizowanej w ramach wykonywania algorytmu. Operacją tą jest tutaj w obydwu przypadkach dokonanie porównania miarą złożoną dwóch rekordów bibliograficznych. Na tak określoną złożoność obliczeniową nie będzie miał wpływu rodzaj realizowanego algorytmu/algorytmów zastosowanych do porównania dwóch rekordów, gdyż wraz z napływem nowych informacji do bazy danych rośnie jedynie liczba jej rekordów, a nie ich długość. Tak więc złożoność obliczeniowa wyliczenia miar podobieństwa względem zadanego opisu bibliograficznego jest liniowa, czyli przyrost zapotrzebowania na moc obliczeniową jest wprost proporcjonalny do przyrostu danych. Do realizacji porównania każdego rekordu z każdym przyrost ten opisywany jest już zależnością bardziej złożoną. Ponieważ operacja wyliczenia miary podobieństwa jest przemienna i nie ma sensu wyliczać tych miar raz dla pary rekordów (a, b) i kolejny raz dla (b, a) oraz wyliczać podobieństwo danego rekordu do samego siebie (a, a), miara ta wyraża się poniższym wzorem:

$$\frac{n^2}{2} - \frac{n}{2}$$

Ponieważ w powyższej formule najwyższy stopień potęgi wynosi 2, to w tym przypadku złożoność obliczeniową musimy określić jako kwadratową.

Warto jednak zwrócić uwagę, że do wdrożenia metody reaktywnej nie jest konieczne wielorazowe przeprowadzanie porównania każdego rekordu z każdym. Wystarczy tylko odnotowywać nowo wprowadzane rekordy i tylko je poddawać weryfikacji z wszystkimi pozostałymi. Podejście takie pozwala sprowadzić to zagadnienie do złożoności liniowej, gdyż liczba dziennie wprowadzanych rekordów oscyluje wokół pewnej stałej wartości.

PODSUMOWANIE

W artykule przedstawiono koncepcję deduplikacji rekordów w bibliograficznych bazach danych. Proces ten nie jest standardową operacją, której wdrożenie w każdym środowisku przebiega identycznie, ale wymaga wcześniejszych analiz typów nieprawidłowości występujących w konkretnej bibliograficznej bazie danych, gdyż zarówno typy te, jak i liczba rekordów dotkniętych nieprawidłowościami danych typów, zależne są zarówno od przyjętego modelu danych, jak i konstrukcji oprogramowania, które je zasila. Przedstawione koncepcje wykrywania duplikatów za pomocą stosowania miar ostrych oraz miar podobieństw łańcuchów znakowych, tak prostych, jak i złożonych, wykorzystujące implementacje poszczególnych typów miar (na przykładzie miary Jaro-Winkler oraz znormalizowanej miary odległości edycji) oraz strojenia miar złożonych, mogą zostać zastosowane do zaprojektowania procesu deduplikacji i jego implementacji w różnych środowiskach bibliograficznych baz danych, zarówno w modelu proaktywnym, jak i reaktywnym.

BIBLIOGRAFIA

Cohen, William W.; Ravikumar, Pradeep; Fienberg, Stephen E. (2003). A comparison of string distance metrics for name-matching tasks. *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI) 18, Workshop on Information Integration on the Web* [online], [dostęp: 04.06.2017]. Dostępny w WWW: <http://www.cs.utexas.edu/users/ai-lab/pubs/ravikumarIJCAI03.pdf>.

Drabik, Adrian (2016). Wyszukiwanie powielonych opisów bibliograficznych w bazie danych: przykład Repozytorium Uniwersytetu Jagiellońskiego. *Przegląd Biblioteczny*, z. 1, s. 65-79.

Dressler, Kevin; Ngonga Ngomo, Axel-Cyrille (2017). On the efficient execution of bounded Jaro-Winker Distances. *Semantic Web*, vol. 8, no. 2, pp. 185-196.

Freire, Nuno; Borbinha, José; Calado, Pável (2007). Identification of FRBR Works Within Bibliographic Databases: An Experiment with UNIMARC and Duplicate Detection Techniques. *International Conference on Asian Digital Libraries (ICADL 2007)* [online], [dostęp: 20.09.2017]. Dostępny w WWW: <https://ai2-s2-pdfs.s3.amazonaws.com/3d87/d4b223c86b21a709705142fd11275e7f04a4.pdf>.

Gu, Lifang; Baxter, Rohan; Vickers, Deanne; Rainsford, Chris (2003). *Record Linkage: Current Practice and Future Directions* [online]. CSIRO Mathematical and Information Sciences; [dostęp: 04.06.2017]. Dostępny w WWW: <http://dc-pubs.dbs.uni-leipzig.de/files/Gu-2003RecordlinkageCurrentpracticeandfuturedirections.pdf>.

Hamming, Richard W. (1950). Error detecting and error correcting codes. *The Bell System Technical Journal*, vol. 29, no. 2, pp. 147-160.

Jaro, Matthew A. (1989). Advances in record-linkage methodology as applied to the 1985 census of Tampa Florida. *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 414-420.

Jiang, Yu; Lin, Can; Meng, Weiyi; Yu, Clement; Cohen, Aaron M.; Smalheiser, Neil R. (2014). Rule-based deduplication of article records from bibliographic databases. *Database: The Journal of Biological Databases and Curation* [online], Jan 16 [dostęp: 04.06.2017]. Dostępny w WWW: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3893659/>.

Kamińska, Anna Małgorzata (2017a). O rozwoju graficznych języków komunikacji. *Zagadnienia Informacji Naukowej*, nr 2 (110).

Kamińska, Anna Małgorzata (2017b). Od druków źródłowych po mapy nauki. Bibliograficzna baza danych GRUBA. W: *Wizualizacja informacji w humanistyce*. Pod red. Małgorzaty Kowalskiej i Veslavy Osińskiej. Toruń: Wydaw. Naukowe Uniwersytetu Mikołaja Kopernika w Toruniu.

Kamińska, Anna Małgorzata (2017c). Potencjał bibliometryczny bibliograficznej bazy danych CYTBIN w świetle prostych i złożonych wskaźników analitycznych. *Bibliotheca Nostra* (w druku).

Kamińska, Anna Małgorzata (2017d). ProBIT – prospektywna metoda tworzenia trawersowalnych indeksów cytowań a współczesne problemy organizacji przestrzeni informacji w tradycyjnych bibliograficznych bazach danych. *Zagadnienia Informacji Naukowej*, nr 1 (109), s. 66-82.

Kamińska, Anna Małgorzata (2017e). Wizualizacje wybranych wskaźników bibliometrycznych na przykładzie bibliograficznej bazy danych CYTBIN. *Toruńskie Studia Bibliologiczne*, nr 2 (19).

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, vol. 14, iss. 1, pp. 10-25.

Левенштейн, В.И. (1965). *Двоичные коды с исправлением выпадений, вставок и замещений символов. Доклады Академии Наукк СССР*, Т. 163, но. 4, с. 845-848.

Marszakowa-Szajkiewicz, Irena (2009). *Badania ilościowe nauki. Podejście bibliometryczne i webometryczne*. Poznań: Uniwersytet im. Adama Mickiewicza.

Wit, Ernst-Jan C.; Gillette, Marie (1999). *What is Linguistic Redundancy? Technical Report* [online]. The University of Chicago; [dostęp: 04.06.2017]. Dostępny w WWW: <http://www.math.rug.nl/~ernst/linguistics/redundancy3.pdf>.

Artykuł w wersji poprawionej wpłynął do Redakcji 25 października 2017 r.

ANNA MAŁGORZATA KAMIŃSKA
Institute of Information and Library Science
Silesia University
e-mail: anna.kaminska@us.edu.pl

STRING SIMILARITY METRICS AND DEDUPLICATION OF RECORDS IN BIBLIOGRAPHIC DATABASES

KEYWORDS: Bibliographic databases. Deduplication of records. String similarity. Records linkage.

ABSTRACT: **Thesis/Objective** – The article presents the method of deduplicating/linking bibliographic records in databases based on the string similarity metrics. The proposal is based on the author's own experience acquired while building a bibliographic database and conducting bibliometric research based on data acquired from publicly available bibliographic databases. The formal description of the method is illustrated with data obtained from the CYTBIN database. **Research methods** – The development of the method required a review of information architecture of selected Polish bibliographic databases and an identification of problems that affect them, resulting not only from data models but also from the construction of their graphical user interfaces. Several string similarity metrics were analyzed and some of them were used as components of the finally proposed compound method. The method enables the evaluation of bibliographic record similarity based on their attributes. **Results** – The results presented on the example of data acquired from CYTBIN database enabled the empirical verification of the proposed method. In addition, the author performed the analysis of the similarity distribution of bibliographic records from the CYTBIN database calculated for the proposed method and for Jaro-Winkler algorithm based on the titles of bibliographic units. **Conclusions** – The proposed method, after adjusting its parameters to the specificity of selected bibliographic databases, can be used to improve the quality of bibliographic data. Depending on the performance of the computer system, the proactive model (the verification before adding a given record to a database) or/and reactive model (the verification of all or just recently added records, performed for instance during a minor system load at daily intervals) can be implemented.