Warszawa, 19.11.2024

dr hab. Agnieszka Mykowiecka
Instytut Podstaw Informatyki PAN
Jana Kazimierza 5, Warszawa

Recenzja rozprawy doktorskiej mgr. inż. Artura Niewiarowskiego

Zastosowanie algorytmu odległości edycyjnej do ilościowej analizy danych tekstowych

Recenzja rozprawy doktorskiej mgr. inż. Artura Niewiarowskiego, zrealizowanej pod opieką dr hab. inż. Marka Stanuszka, wykonana została na zlecenie Rady Naukowej Instytutu Podstawowych Problemów Nauki i Techniki PAN.

Zawartość pracy

Recenzowana rozprawa doktorska opisuje wyniki prac Doktoranta dotyczących podobieństwa tekstów i składa się z 4 rozdziałów. Po wstępie, w rozdziale 2 zatytułowanym „Przegląd wybranych algorytmów analizy danych tekstowych", autor przedstawia krótkie charakterystyki wybranych metod rozwiązywania niektórych problemów związanych z przetwarzaniem danych tekstowych. Pierwszy podrozdział „Grupowanie dokumentów a ważenie terminów w modelu wektorowym" rozpoczyna przedstawienie wektorowej reprezentacji dokumentu i współczynnika tf-idf. W kolejnym, scharakteryzowanych jest kilka algorytmów stemmingu oraz algorytm soundex. Następny podrozdział poświęcony jest metodom analizy porównawczej zbiorów tekstowych. Autor opisuje tu parę metod wyszukiwania wzorców, a następnie współczynniki Dice'a i Jaccarda oraz miarę cosinusową pozwalające na określenie stopnia podobieństwa tekstów. Następnie przedstawiona jest klasyczna definicja algorytmu Levensteina pozwalającego na efektywne znalezienie odległości edycyjnej między napisami. Rozdział kończy charakterystyka implementacji algorytmów wyszukiwania w tekstowych bazach danych na przykładzie MariaDB.

Zasadniczą część rozprawy rozpoczyna rozdział 3 zawierający opis wprowadzonych modyfikacji algorytmu wyznaczającego odległość edycyjną napisów tak, by służył do wyznaczania podobieństwa zbiorów tekstowych. Autor wprowadza tu miarę podobieństwa, w której poza odległością Levensteina uwzględnia długość dłuższego z porównywanych napisów. Daje to naturalne przeskalowanie wagi liczby różnic w tekście w stosunku do całej jego długości. Na potrzeby porównywana tekstów, autor traktuje słowa jako odpowiedniki znaków w oryginalnym algorytmie i wprowadza miarę podobieństwa wykorzystującą odległość Levensteina dla poszczególnych elementów zdania. Określany dla tego podobieństwa próg decyduje o tym czy uznajemy słowa za jednakowe, czy różne. Autor dobiera wartość tego progu eksperymentalnie, w oparciu o obserwacje dotyczące konkretnej dziedziny lub języka. Kolejnym rozszerzeniem algorytmu jest zaproponowanie użycia słownika wyrazów bliskoznacznych (np. samochód, auto) lub semantycznie do pewnego stopnia równoważnych (jak na przykład różne imiona lub owoce). Doktorant przeprowadził wiele eksperymentów wykazujących zwiększenie współczynnika podobieństwa tekstów. W szczególności testy przeprowadzane były na parach zdań, w których drugie zdanie uzyskane było po tłumaczeniu na i z niemieckiego, angielskiego i rosyjskiego. Teksty, z założenia bardzo podobne, mają wysoką wartość współczynnika podobieństwa, przy czym, zgodnie z celem autora, wprowadzenie podobieństwa słów zwiększa podobieństwo całych tekstów w sposób istotny. Autor dokonał też porównania czasu wykonywania obliczeń dla algorytmu z dodanymi opcjami liczenia podobieństwa i bez nich (przy

założeniu kwadratowej złożoności algorytmu podstawowego) i wskazał na znaczny koszt wyliczania podobieństwa wszystkich słów do pozostałych. O ile w przypadku krótkich zdań nie jest to cecha bardzo uciążliwa, przy dłuższych tekstach koszt algorytmu staje się nieakceptowalny. Zaproponowanym przez doktoranta rozwiązaniem jest analiza macierzowa polegająca na ocenie macierzy podobieństwa słów i sterowaniu algorytmem oceniającym poprzez ustawione odpowiednio wartości parametrów mówiących na jakie odstępstwa pozwalamy (na jakim poziomie akceptujemy podobieństwo słów, jaką przerwę między podobnymi słowami akceptujemy i jak długich podobnych fragmentów oczekujemy). Propozycja ta, łączenie z przedstawieniem wyników licznych testów, opisana jest w rozdziale czwartym – Weryfikacja przedstawionego mechanizmu analizy danych tekstowych.

Piąty rozdział zawiera podsumowanie, będące częściowo powtórzeniem podrozdziału 4.9 (Wnioski). Dodatkowo przedstawiony jest tu opracowany przez doktoranta program do wykrywania plagiatów – Antyplagius – opracowany w ramach działalności firmy New Data Mining Systems sp. z.o.o. Program ma ten ma też dodatkowe funkcjonalności pozwalające na przeszukiwanie zawartości dysku w wykorzystaniem zaproponowanej w pracy metody ustalania podobieństwa napisów.

Ocena

Praca doktorska Artura Niewiarowskiego przygotowywana była dość długo (od 2017 roku). Długi okres przygotowywania pracy z jednej strony pozwala na pogłębione badania, z drugiej niesie ryzyko utraty aktualności prowadzonych prac. Ten drugi aspekt jest w tym przypadku istotny, gdyż w ostatnim czasie w dziedzinie przetwarzania tekstów w języku naturalnym nastąpiła rewolucja technologiczna. Masowe zastosowanie metod maszynowego uczenia się i powstanie nowych modeli języka pozwoliło na uzyskiwanie rezultatów nieosiągalnych tej pory innymi metodami. Mgr Niewiarowski zauważył oczywiście istnienie tych nowych metod, wykorzystał modele generatywne i współczesne programy do maszynowego tłumaczenia do wygenerowania danych tekstowych. Nie porównał jednak swoich prac do jakiejkolwiek próby ustalenia podobieństwa tekstów za pomocą innych metod. Oczywiście rozumiem założenie autora, który z definicji nie chciał z tych metod korzystać, ale choćby spekulacyjne rozważania o tym jak to wygląda w obecnym kontekście, byłyby pożądane. Niezależnie od tego, wskazanie skutecznej, alternatywnej metody wykazującej się dobrymi efektami i nie wymagającej dużych obcych zasobów danych czy gotowych modeli, jest rozwiązaniem dla wielu potencjalnie atrakcyjnym. Zaprezentowana metoda dając wizualny obraz podobieństwa pozwala na precyzyjne wskazanie fragmentów podobnych. Oczywiście w przypadku dłuższych tekstów oglądanie takiej macierzy jest utrudnione, ale zaproponowane metody analizy podciągów wybranej długości może to ułatwić. Metoda macierzowa ustalania podobieństwa jest przedstawiona w dużej mierze jako metoda graficzna. Autor przedstawia też ogólny liczbowy wskaźnik podobieństwa, ale nigdzie nie znalazłam sformułowanego explicite wzoru na jego wyliczenie. Zaprezentowana metoda skupia się na podobieństwie ortograficznym, autor nie uwzględnia zatem możliwych różnic wynikających z wieloznaczności słów, czy różnego znaczenia słów edycyjnie do siebie bardzo podobnych (np. sad-sąd). Jednak w przypadku nastawienia na znajdowanie podobieństw nie w krótkich frazach, ale zdaniach czy w tekstach, przykładowo w celu wykrywania plagiatów, te niezgodności nie wydają się bardzo istotne. Autor zauważa natomiast ważny problem nieuwzględnienia w jego algorytmie możliwości zamiany porządku elementów tekstu. Brakuje jednak analizy, czy zwiększanie progu dla miary podobieństwa dla podobnych tekstów nie wpływa negatywnie na ocenę tekstów niepodobnych.

Przechodząc do uwag bardziej szczegółowych, zastrzeżenia moje budzi wprowadzająca część pracy, która opisuje problem w sposób niepełny. Opisane w rozdziale drugim metody wydają się być wybrane w sposób nieco przypadkowy i odzwierciedlają niewielki zakres tema-

tyczny. Pewnym zgrzytem już na początku jest błąd przy opisie współczynnika tf-idf, który nie mógł wynikać z przypadkowej pomyłki. Poświęcenie w tym samym rozdziale relatywnie dużo miejsca algorytmom stemmingu wydaje się mało uzasadnione. Autor wielokrotnie podkreśla, że tych algorytmów nie używa, gdyż jego metoda ma je niejako zastąpić i nie porównuje wyników swojej metody z taką wykorzystującą stemmer. Narzędzia te, obecnie nie są już chyba praktycznie używane. W pracy Doktorant często pisze łącznie stemming i lematyzacja, nie rozdzielając tych metod, dających inne wyniki. Sam jako wynik takiej transformacji przedstawia efekt wychodzący poza to, co rozumiemy poprzez lematyzację. W szczególności w tabeli 3, forma *ma* zmieniona jest na na *posiadać*, czego żaden ze znanych mi algorytmów stemmingu ani lematyzacji nie uczyniłby. Opisany następnie algorytm Soundex, rozwiązanie polegające na kodowaniu w identyczny sposób słów, które pisane są nieco inaczej, a brzmią tak samo (lub bardzo podobnie), nie był wykorzystywany przy prowadzonych pracach. Także ciekawy skądinąd problem wyszukiwania w bazach tekstowych nie został wprowadzony tak, by widać było wyraźniej jego związek z tematem. Autor pisze w pracy o możliwości użycia słownika wyrazów bliskoznacznych, gdzieś wspomina o plWordent, ale nie wiąże tych tematów i nie wskazuje czy sam jakiegoś słownika używał. W sposób niezbyt uporządkowany wprowadzona jest miara odległości cosinusowej i nie do końca wiadomo jak budowane są w tym przypadku wektory, dla których jest wyliczana. Zasadnicza część pracy napisana jest w sposób uporządkowany i przejrzysty, ciąg wprowadzania poszczególnych pojęć i zaproponowanych eksperymentów jest właściwy, a poziom szczegółowości duży (czasem chyba aż za duży w miejscach tego nie wymagających). Z drobnych uwag - na stronie 53 przedstawiony jest algorytm wypełniania macierzy liczbami 0 i 1, który jest raczej zbyteczny, bo oczywisty, warto byłoby jednak napisać co znaczy tu 'term'. Gdzieś dalej w tekście wspomniane jest o tokenizacji na spacjach i pomijaniu słów jednoliterowych, ale nie wiadomo, czy tak jest zawsze. Na stronie 57 brakuje sformułowania jak liczony jest końcowy wynik porównania. Zestaw przedstawionych eksperymentów jest obszerny i wskazuje na praktyczną użyteczność zaproponowanego rozwiązania. Sam tekst pracy jest płynny i poprawny, błędy językowe zdarzają się bardzo sporadycznie. Wartościowym efektem dodatkowym prowadzonych badań jest na pewno zaimplementowany program, który potencjalnie może mieć różne praktyczne zastosowania.

Wniosek końcowy

Przedłożona mi do recenzji rozprawa zawiera opis autorskiej metody porównywania tekstów. Jest to stosunkowo prosta metoda wykorzystująca zmodyfikowaną metodę liczenia odległości Levensteina między napisami. Mimo prostoty metoda wykazała swoją skuteczność w wielu różnie zaprojektowanych eksperymentach sprawdzających podobieństwo tekstów w tym samym i różnych językach i różnych alfabetach. Częściowe wyniki prac Doktoranta opublikowane zostały w kilku artykułach w czasopismach i materiałach konferencyjnych. Sama praca pozostawia pewien niedosyt, gdyż część zagadnień nie zostało w niej ukazanych w szerszym naukowym kontekście, a wysiłek doktoranta skierowany był raczej na wykonanie większej liczby eksperymentów. Może to być częściowo wynik tego, że Doktorant prowadząc zajęcia na wydziale technicznym nie pracował raczej w środowisku osób zajmujących się analizą tekstów oraz być może pewnego niedostatku opieki promotora. Mimo tych zastrzeżeń uważam, że opracowanie, zaprezentowanie i przetestowanie własnej metody postawionego problemu jest osiągnięciem, które może stanowić podstawę do nadania stopnia doktora i wnoszę o dopuszczenie magistra inż. Artura Niewiarowskiego do publicznej obrony.

Podpisane elektronicznie przez Agnieszka Maria Mykowiecka (Certyfikat kwalifikowany) w dniu 2024-11-21.

Agnieszka Mykowiecka