Streszczenie rozprawy doktorskiej: „Automatyczna klasyfikacja dokumentów tekstowych w zarządzaniu aktami spraw na użytek e-administracji"

mgr Marek Gajewski

Sformułowanie zadania

Inspiracją podejmowanego w rozprawie zadania jest zagadnienie zarządzania *aktami spraw* w ramach obiegu dokumentów. W prawodawstwie polskim wymagane jest, aby dokumentacja gromadzona przez jednostki administracji publicznej i sposób jej organizacji odzwierciedlały przebieg załatwiania *sprawy*. Organizacja i klasyfikacja dokumentów prowadzona jest na podstawie systematyzacji rzeczowej *akt spraw* znanej pod pojęciem *Jednolitego Rzeczowego Wykazu Akt* (JRWA) i stanowiącej hierarchię klas, do których powinny być zaklasyfikowane wszystkie przetwarzane dokumenty. W ogólności oznacza to, że dla każdego nowego dokumentu należy rozpoznać jego przynależność do klasy z JRWA oraz w ramach tejże klasy przypisać go do odpowiednich *akt sprawy*. Tak sformułowany problem decyzyjny stanowi inspirację dla zadania postawionego w niniejszej rozprawie – automatycznego przypisywania dokumentu do klasy i pewnej sekwencji dokumentów (sprawy) w ramach tej klasy.

Zadanie postawione w niniejszej pracy odnosi się więc do problemu dwóch rodzajów klasyfikacji dokumentu, rozpatrywanych łącznie: do tradycyjnie rozumianej klasy (kategorii) dokumentów (w praktycznym scenariuszu stanowi ją klasa JRWA) i do konkretnej sprawy, wyodrębnionej w ramach tej klasy. Z jednej strony należy zaklasyfikować dokument zgodnie z jego *tematyką*, z drugiej zaś strony, zgodnie z, szeroko rozumianą, *sprawą*, której on dotyczy. Pierwsze z zadań ma bogatą literaturę, liczne zaproponowane w niej sposoby rozwiązania, i często jest rozważane pod nazwą *kategoryzacja dokumentów tekstowych* (ang. *text categorization*). Drugi z rozważanych rodzajów klasyfikacji jest bardziej złożony i wymaga przypisania do tej samej klasy (sprawy) dokumentów, które są ze sobą powiązane tym, że dotyczą tego samego zdarzenia, zamówienia, osoby itp. To powiązanie ma charakter wykraczający poza odniesienie się wyłączne do tematyki dokumentów. Co więcej, zbiór klas jest w tym wypadku dynamiczny: dokument może należeć do istniejącej już klasy (sprawy) lub może stanowić początek nowej sprawy i wymaga utworzenia nowej klasy. Kolejną trudność stanowi fakt, że sprawy (klasy) obejmują przeciętnie niezbyt wiele dokumentów, a więc brak jest dostatecznej liczby dokumentów uczących dla skutecznego zastosowania znanych algorytmów klasyfikacji. Ten dwojaki charakter klasyfikacji stanowi o oryginalności postawionego zadania.

Formalnie, zadanie postawione w rozprawie można przedstawić następująco. Przyjmijmy w tym celu następującą terminologię i notację:

$D = \{d_1, \ldots, d_n\}$ - korpus dokumentów, z których każdy przypisany jest do jednej z klas (kategorii) i występuje w jednej ze spraw (sekwencji dokumentów) związanych z tą klasą,

d^* - dokument zadany na wejściu, który należy przypisać do jednej z kategorii i jednej ze spraw należących do tej kategorii,

$C = \{c_1, \ldots, c_m\}$ - zbiór rozważanych klas (kategorii) dokumentów,

$\Sigma = \{\sigma_1, \ldots, \sigma_p\}$ - zbiór spraw (sekwencji dokumentów), każda należąca do jednej z klas c_i,

$\sigma_k = < d_{s_1}, d_{s_2}, \ldots, d_{s_l} >$, sprawa (sekwencja dokumentów), gdzie wszystkie dokumenty d_i należą do jednej klasy c_j; wyróżniamy sprawy „w toku", do których należy klasyfikować nowy dokument oraz sprawy już „zamknięte".

Ostatecznym celem w rozważanym zadaniu jest zaklasyfikowanie dokumentu d^* do jednej ze spraw „w toku" σ_k lub do nowej sprawy, a przez to również zaklasyfikowanie do jednej z kategorii c_j.

Metoda rozwiązania

Reprezentacja danych

Dokument

W zastosowanym w niniejszej pracy podejściu, jako punkt wyjścia do opracowania reprezentacji dokumentów przyjęto klasyczny model wektorowy. Każdy z dokumentów $d \in D$ jest reprezentowany za pomocą wektora $[w_1, \ldots, w_M]$ w przestrzeni $T = \{t_i\}_{i \in \{1,\ldots,M\}}$, gdzie T jest zbiorem wszystkich słów kluczowych używanych do indeksowania dokumentów w kolekcji, przy czym ostatecznie używa się tylko K słów kluczowych, odpowiadających współrzędnym o największych wartościach w tym wektorze. Wartość parametru K przyjmuje się dostatecznie małą, na przykład równą 5, w celu uzyskania bardziej zwartej reprezentacji dokumentów.

Formalnie reprezentacja dokumentu jest następująca:

$$d \longrightarrow \left[w_1^d, \ldots, w_M^d\right] \rightarrow \left[w_{k_1}^d, \ldots, w_{k_K}^d\right] \tag{1}$$

gdzie w_i^d oznacza stopień ważności słowa kluczowego t_i dla reprezentacji dokumentu $d \in D$, zaś $k_1, \ldots, k_K$ są indeksami takimi, że:

$$w_{k_1}^d \geq w_{k_2}^d \geq \ldots \geq w_{k_K}^d \geq w_{k_{K+1}}^d \geq \cdots \geq w_{k_M}^d$$

Sprawa (sekwencja dokumentów)

Sekwencje dokumentów $\sigma = \langle d_{s_1}, \ldots, d_{s_l} \rangle$ tworzących *sprawy* reprezentowane są przez połączenie wektorów reprezentujących kolejne składowe dokumenty. Formalnie:

$$\sigma \longrightarrow \left[w_{k_1}^{d_{s_1}}, \ldots, w_{k_K}^{d_{s_1}}, \ldots, w_{k_1}^{d_{s_l}}, \ldots, w_{k_K}^{d_{s_l}}\right] \tag{2}$$

Przy obliczaniu dopasowania dokumentu $d^* \longrightarrow \left[w_{k_1}^{d^*}, \ldots, w_{k_K}^{d^*}\right]$ do sekwencji (2), jej reprezentacja ulega zmianie i przyjmuje następującą postać:

$$\sigma \longrightarrow \left[w_{k_1}^{\sigma}, \ldots, w_{k_K}^{\sigma}\right] \tag{3}$$

gdzie k_i oznacza indeks i-tego słowa kluczowego, t_{k_i}, uwzględnionego w reprezentacji dokumentu d^*, $i \in [1, K]$, zaś jego stopień ważności $w_{k_i}^{\sigma}$ w reprezentacji sekwencji (2) jest obliczany następująco:

$$w_{k_i} = \begin{cases} 0 & \text{jeśli } t_{k_i} \text{ nie występuje w reprezentacji żadnego z dokumentów } \sigma \\ w_{k_j}^{d_{sm}} & \text{w przeciwnym przypadku} \end{cases} \tag{4}$$

gdzie $k_j = k_i$ oraz $m \in [1, l]$ jest największym indeksem dokumentu w sekwencji σ (indeksem dokumentu występującego w sekwencji najpóźniej), w którego reprezentacji występuje słowo k_i.

Kategoria

Reprezentacja kategorii $c \in C$ przybiera postać średniej wszystkich wektorów reprezentujących poszczególne dokumenty aktualnie przypisane do kategorii c. Formalnie:

$$c \rightarrow \left[w_1^c, \ldots, w_{M_c}^c\right] \tag{5}$$

gdzie

$$w_{t_i}^c = \frac{\sum_{d \in c} w_{t_i}^d}{|\{d : d \in c\}|} \tag{6}$$

gdzie $d \in c$ oznacza, że dokument d przypisany jest do kategorii c, a $|\cdot|$ oznacza liczność zbioru. Reprezentacja dokumentów używana w (6) jest określona zgodnie z (1). Tak więc, słowo kluczowe t_i pojawi się w reprezentacji kategorii c o ile tylko występuje ono w reprezentacji przynajmniej jednego z dokumentów przypisanych do tej kategorii. We wzorze (6) brane są pod uwagę wszystkie dokumenty znajdujące się aktualnie w kolekcji.

Klasyfikacja

Klasyfikacja dokumentu odbywa się na podstawie oceny jego podobieństwa do poszczególnych spraw i kategorii z nimi związanych. Dla każdego nowego dokumentu d^* oraz każdej sprawy „w toku" σ obliczane są dwa stopnie podobieństwa, odpowiednio:

$$\mathrm{sim}_{seq}(d^*, \sigma) \tag{7}$$

$$sim_{cat}(d^*, c) \tag{8}$$

przy czym pierwszy z nich dotyczy danej sprawy σ, a drugi dotyczy kategorii, do której sprawa ta należy. Następnie, łączne podobieństwo dokumentu d^* do sprawy σ obliczane jest jako suma ważona podobieństw (7) – (8):

$$w_{seq} * sim_{seq}(d^*, \sigma) + w_{cat} * sim_{cat}(d^*, c) \tag{9}$$

Dokument przypisywany jest do tej sprawy, dla której wartość podobieństwa wyliczona z użyciem wzoru (9) jest największa. Ten ogólny schemat algorytmu klasyfikacji jest realizowany poprzez dobór:

1) odpowiednich sposobów obliczania podobieństwa, sim_{seq} i sim_{cat}, oraz
2) stopni ważności, w_{seq} i w_{cat}.

Tak więc mamy tu do czynienia z całą rodziną algorytmów, które uzyskuje się z algorytmu bazowego przez przyjęcie odpowiednich sposobów obliczania podobieństwa i stopni ważności.

W zaproponowanym w rozprawie podejściu, sposób liczenia podobieństwa odwołuje się do pojęcia *zawierania się zbiorów rozmytych*. Zarówno klasyfikowany dokument d^*, jak i reprezentacje spraw σ i kategorii c, traktuje się jako zbiory rozmyte i podobieństwa (7)-(8) oblicza się, odpowiednio, jako stopień zawierania się d^* w σ i w c. To rozwiązanie okazuje się być skuteczne i efektywne.

Rozwiązanie zadania opiera się więc na zastosowaniu wskaźnika zawierania się zbiorów rozmytych, $sub(A, B)$:

$$sub: [0,1]^X \times [0,1]^X \rightarrow [0,1] \tag{10}$$

gdzie X oznacza przestrzeń rozważań.

Określa on stopień, w jakim zbiór rozmyty A jest podzbiorem zbioru rozmytego B. W rozprawie rozważa się zastosowanie wielu różnych wskaźników, w tym wskaźnik zawierania Kosko i wskaźniki bazujące na operatorze implikacji rozmytej i ich warianty.

W zadaniu archiwizacji, stanowiącym inspirację dla zadania rozważanego w rozprawie, występuje dodatkowy aspekt, który został uwzględniony w proponowanym rozwiązaniu. Przyjmuje się mianowicie, że kategorie $c \in C$, do których należy przypisać dokumenty, stanowią liście pewnej struktury hierarchicznej (drzewiastej) H. Kategorie te są więc ze sobą powiązane i to ich powiązanie może zostać wykorzystane w procesie klasyfikacji dokumentów. Zasadnicza idea takiego wykorzystania

hierarchii polega na założeniu, że jeśli dokonuje się wyboru pomiędzy kategoriami (liśćmi hierarchii) posiadającymi tego samego rodzica w hierarchii H, to przy ustalaniu wspomnianego stopnia zawierania się, rola słów kluczowych wspólnych dla tych kategorii jest mniej znacząca i wprowadza się dodatkowy stopień ważności słów kluczowych, który ma tę redukcję znaczenia reprezentować.

W związku z tym, w rozprawie proponuje się rozszerzenie pojęcia zawierania się zbiorów rozmytych z uwzględnieniem stopnia ważności elementów przestrzeni rozważań:

$$sub^{W}: [0,1]^{X} \times [0,1]^{X} \times [0,1]^{X} \rightarrow [0,1] \tag{11}$$

i bada się sposoby zaadoptowania znanych wskaźników zawierania się zbiorów rozmytych, w szczególności wskaźnika Kosko, tak aby uwzględniały one taki dodatkowy stopień ważności.

Eksperymenty obliczeniowe

Sformułowany problem jest nowy i nie istnieją jeszcze standardowe korpusy danych testowych dla niego. W trakcie prac nad rozprawą podjęto próby stworzenia takich korpusów. Wykorzystano w tym celu dwa popularne zbiory danych: kolekcję publikacji z zakresu przetwarzania języka naturalnego i lingwistyki obliczeniowej, znaną jako ACL Anthology Reference Corpus (ACL ARC) oraz zestaw dokumentów znany w literaturze jako korpus Browna. W obydwu wypadkach, dokumenty składające się na poszczególne korpusy utożsamiono ze sprawami, a wydzielone z nich części jako dokumenty składające się na daną sprawę.

Przeprowadzono liczne eksperymenty obliczeniowe z użyciem zaproponowanego algorytmu i różnych konfiguracji jego parametrów (konkretnych wskaźników zawierania się zbiorów rozmytych (7)-(8) i stopni ważności w_{seq} i w_{cat} (we wzorze (9)). Eksperymenty obliczeniowe miały na celu porównanie zaproponowanej metody z bazowym algorytmem klasyfikacji typu *k-nn* oraz zweryfikowanie efektywności różnych parametryzacji zaproponowanego algorytmu.

Wnioski

W rozprawie zaproponowano nowe podejście do analizy i rozwiązania konkretnego problemu kategoryzacji dokumentów tekstowych, w którym każdy dokument, oprócz przynależności do właściwej kategorii, należy do określonej sekwencji dokumentów, rozumianej tutaj jako sprawa, zawierającej dokumenty z tej samej kategorii. Zaproponowany algorytm łączy dwa wskaźniki względem których dokument jest klasyfikowany: jeden odpowiada za podobieństwo do sprawy a drugi odzwierciedla podobieństwo do kategorii. Wskaźniki te są oparte na wskaźnikach zawierania się zbiorów rozmytych. Zbadano skuteczność i efektywność zaproponowanego algorytmu dla różnych kombinacji wag obu wskaźników oraz zastosowanych wskaźników zawierania się zbiorów. Uzyskane wyniki są bardzo zachęcające. Jednak dalsze eksperymenty na różnych zbiorach dokumentów są potrzebne, aby możliwe było wyciągnięcie dalej idących wniosków. Otrzymane wyniki pokazują, że zaproponowana dość prosta metoda może być skuteczna i wydajna w radzeniu sobie z trudnym problemem klasyfikacji dokumentów w przypadku niewielkiej reprezentacji klas (spraw).

Ważnym założeniem było zastosowanie niskowymiarowej reprezentacji dokumentów i innych typów danych niezbędnych do realizacji zadania. Osiągnięto to przez przyjęcie wektorowej reprezentacji dokumentów z agresywną redukcją liczby uwzględnianych słów kluczowych.

Wprowadzono również pojęcie zawierania się zbiorów rozmytych z uwzględnieniem stopni ważności poszczególnych elementów przestrzeni rozważań. Pojęcie to wynikło w trakcie rozwiązywania zadania postawionego w rozprawie, a dokładniej jego wariantu, w którym dla rozróżnienia pokrewnych hierarchicznie kategorii lub spraw należących do tej samej kategorii

interesujące może być uwzględnienie zróżnicowanego znaczenia poszczególnych słów kluczowych przy klasyfikowaniu dokumentu. Pojęcie to ma jednak ogólniejszy charakter i jest interesujące samo w sobie. W rozprawie zaproponowano rozszerzenie poszczególnych wskaźników zawierania się zbiorów o uwzględnienie takich dodatkowych stopni ważności poszczególnych elementów. W szczególności przeprowadzono pogłębioną analizę takiego rozszerzenia w wypadku wskaźnika Kosko. Pokazano, że naiwne rozszerzenie tego wskaźnika może prowadzić do teoretycznie nieintuicyjnego jego zachowania, a użycie innej interpretacji bazowego wskaźnika jako punktu wyjścia pozwala przezwyciężyć tę trudność.

Marek Gajewski