STUDIA INFORMATICA 2009
Volume 30 Number 2A (83)

Piotr ANDRUSZKIEWICZ
Politechnika Warszawska, Instytut Informatyki

REPREZENTACJE DOKUMENTÓW TEKSTOWYCH W KONTEKŚCIE WYKRYWANIA NIECHCIANYCH WIADOMOŚCI POCZTOWYCH W JĘZYKU POLSKIM Z WTRĄCENIAMI W JĘZYKU ANGIELSKIM

Streszczenie. Klasyfikacja dokumentów tekstowych wiąże się z utworzeniem ich reprezentacji. Duża liczba dokumentów zachęca do prób stosowania jak najbardziej oszczędnych sposobów ich reprezentowania. W niniejszej pracy przedstawione zostały możliwe reprezentacje dokumentów tekstowych, sposoby ich ograniczania w kontekście wykrywania niechcianych wiadomości pocztowych w języku polskim z wtrąceniami w języku angielskim.

Słowa kluczowe: reprezentacja dokumentów tekstowych, funkcje istotności atrybutów, TF-IDF, ograniczanie reprezentacji, klasyfikacja, wykrywanie niechcianej poczty elektronicznej

REPRESENTATIONS OF TEXT DOCUMENTS IN CONTEXT OF SPAM DETECTION IN POLISH WITH ENGLISH PHRASES

Summary. Representation of text documents should be as small as possible and give high accuracy of classification. This paper presents representations of text documents and ways of their reduction in case of SPAM detection in Polish with English phrases.

Keywords: text document representation, term weighting functions, TF-IDF, reduction of text document representation, classification, SPAM detection

1. Wstęp

W celu przeprowadzenia klasyfikacji dokumentów tekstowych należy stworzyć ich reprezentację. Duże wolumeny analizowanych tekstów i ich liczba powodują, że celowe są

próby wykorzystywania oszczędnych sposobów reprezentacji. Zbyt mocne ograniczanie informacji na temat analizowanych dokumentów prowadzi do zmniejszenia jakości klasyfikacji. Pojawia się więc problem doboru właściwej reprezentacji, która jednocześnie zapewni wysoką jakość klasyfikacji (zbliżoną do sytuacji bez zmniejszania rozmiaru reprezentacji) oraz pozwoli na ograniczenie przestrzeni zajmowanej przez analizowany zbiór dokumentów i ograniczenie czasu przetwarzania.

Innym problemem pojawiającym się przy wykrywaniu niepożądanych wiadomości w poczcie elektronicznej w języku polskim jest występowanie w takich dokumentach także zwrotów w innych językach. Najczęściej są to wyrażenia w języku angielskim, który jest „językiem świata”[1] – stosowany jest w publikacjach naukowych, programowaniu, a dodatkowo jest językiem biznesu.

Przedstawiony problem „zanieczyszczenia” języka powoduje, że konieczne jest specyficzne podejście do tworzenia reprezentacji dokumentów. Możliwe rozwiązanie przedstawione zostało w niniejszej pracy.

W pracy poruszane są następujące zagadnienia. Rozdział 2 to przegląd literatury. Rozdział 3 przedstawia proponowane w literaturze reprezentacje dokumentów tekstowych, natomiast rozdział 4 opisuje sposoby zmniejszania rozmiarów dokumentów. W rozdziale 5 analizowane jest nowe podejście do klasyfikacji dokumentów tekstowych w języku polskim z wtrąceniami w języku angielskim. Rozdział 6 to wyniki eksperymentów przeprowadzonych przy użyciu stworzonego systemu. Rozdział 7 to podsumowanie i wskazanie kierunków dalszych prac.

2. Przegląd literatury

Eksploracja tekstu jest silnie rozwijającą się dziedziną. Dostępna więc jest bogata literatura dotycząca reprezentacji dokumentów tekstowych [10, 1, 9] oraz funkcji istotności atrybutów [20, 27, 6].

W analizie dokumentów tekstowych można wyróżnić następujące zadania: wyszukiwanie (ang. *Information Retrieval*) [7], klasyfikacja (ang. *classification*) [15], grupowanie (ang. *clustering*) [13, 29], generowanie streszczeń (ang. *text summarization*) [16]. Przy analizie dokumentów mówi się także o eksploracji danych w sieci, Internet (ang. *web content mining, web text mining*) [4, 24], która jest specyficzna ze względu na występowanie formatowania i grafiki w dokumentach.

[1] Język służący komunikowaniu się różnojęzycznych grup nazywany jest także *lingua franca* (wł. język Franków). Rolę taką w przeszłości pełnił starożytny grecki koine, łacina czy język francuski oraz niemiecki.

Przegląd rozwiązań stosowanych w tworzeniu reprezentacji dokumentów tekstowych oraz ich klasyfikacji różnymi algorytmami można znaleźć w [5]. Praca ta nie odnosi się jednak do specyfiki dokumentów tekstowych w języku polskim. Nie przedstawia również praktycznych eksperymentów i porównań.

W literaturze dostępne są także opracowania, które skupiają się na klasyfikacji dokumentów w języku polskim [3], który znacząco różni się od języka angielskiego przez znacznie bardziej złożoną gramatykę oraz większą fleksyjność. Można znaleźć także opracowania dotyczące wykrywania niechcianej poczty elektronicznej pisanej tylko w języku angielskim [25, 8, 31, 30] lub tylko w języku polskim [28].

Nie poruszano w literaturze jednak kwestii klasyfikacji dokumentów, w szczególności wiadomości poczty elektronicznej, w języku polskim z wtrąceniami w języku obcym, najczęściej angielskim. Problem ten przedstawiony został w niniejszej pracy.

3. Reprezentacje dokumentów

W celu przeprowadzenia eksploracji tekstu (ang. *text mining*) należy wykonać transformację dokumentów do postaci akceptowanej przez algorytm, który zostanie użyty. Stosowane reprezentacje dokumentów tekstowych wykorzystywane w klasyfikacji, ale także i grupowaniu przedstawione zostały poniżej.

3.1. Reprezentacje unigramowe – binarna i częstościowa

Najprostszą reprezentacją, ale i często stosowaną, jest reprezentacja unigramowa [26]. Polega ona na zliczaniu częstości występowania słów w treści dokumentu. Częstości te wykorzystuje się do zbudowania wektora reprezentującego cały dokument. Zliczanie to można przeprowadzić na dwa sposoby, odnotowując jedynie fakt wystąpienia słowa w danym dokumencie (reprezentacja binarna) lub zapisując liczbę wystąpień (reprezentacja częstościowa). Druga reprezentacja jest bardzo wrażliwa na długość dokumentu, co jest zjawiskiem niepożądanym. Wprowadza się więc częstotliwość względną, czyli odniesioną do liczby wszystkich słów w dokumencie.

3.2. Reprezentacja n-gramowa

Reprezentacja ta jest bardziej rozbudowaną reprezentacją, niosącą ze sobą więcej informacji. Uwzględnia ona bowiem szyk słów w zdaniu. Przechowuje informacje o następstwie wyrazów. Zakłada się, że to, jakie słowo występuje na k-tej pozycji, zależy od n poprzedzają-

cych je wyrazów. Założenie takie nie jest prawdziwe, mimo to jest lepszym przybliżeniem rzeczywistego języka [23].

3.3. Reprezentacja γ-gramowa

Polega na przechowywaniu informacji o częstości wystąpień sekwencji wyrazów o różnej długości [10]. Budowę takiej reprezentacji zaczyna się od zliczenia słów pojedynczych. Na ich podstawie tworzy się zbiór podwójnych słów, następnie potrójnych itd. Dzięki takiej reprezentacji nie występuje problem losowania pierwszego wyrazu (dla reprezentacji n-gramowej należało przyjąć jakieś warunki początkowe), wystarczy odnieść się do prawdopodobieństw dla zbioru słów jednowyrazowych.

4. Ograniczanie wielkości reprezentacji

Tworząc reprezentację dokumentów tekstowych, dokonuje się usunięcia szumu informacyjnego oraz ograniczenia reprezentacji. Pierwszy element ma na celu poprawienie skuteczności klasyfikacji przez usunięcie zbędnych słów, które mogą negatywnie wpływać na proces klasyfikacji. Realizuje się go w przetwarzaniu wstępnym. Powoduje on także ograniczenie objętości reprezentacji. Realizacja drugiego zadania wykonywana jest zazwyczaj przez wybór cech – termów (ang. *feature selection*), które najlepiej charakteryzują dokumenty. Celem jest uzyskanie jak najmniejszej liczby cech, które pozwolą na zachowanie wymaganego poziomu jakości klasyfikacji. Poniżej opisano szczegółowo poszczególne pojęcia.

4.1. Przetwarzanie wstępne

Proces ten realizowany jest przed stworzeniem reprezentacji. Można wyróżnić jego dwie główne części.

4.1.1. Zastosowanie stop-listy

Wśród słów analizowanego tekstu znajdują się wyrazy, które nie niosą informacji. Są to na przykład: spójniki, rodzajniki, przyimki itp. Mogą one zostać usunięte bez szkody dla jakości analizy tekstu, a nawet mogą ją poprawić poprzez usunięcie szumu. Lista takich słów nazywana jest stop-listą (ang. *stopword list*) [21]. Należy jednak uważnie stosować tę listę, gdyż usunięcie słów, które wydają się być bez znaczenia dla analizy dokumentu, może spowodować zmianę znaczenia poszczególnych fraz [10].

4.1.2. Sprowadzanie wyrazów do formy podstawowej

Innym sposobem na ograniczenie reprezentacji (zmniejszenie liczby termów), które może także prowadzić do poprawy jakości klasyfikacji, jest sprowadzenie wyrazów do ich formy podstawowej (ang. *stemming*) – wyznaczenie rdzenia dla danego słowa. Wówczas słowa „biegał" i „biegali" zastąpione zostanę przez jeden wyraz. W przypadku eksploracji tekstu nie ma wymogu, aby lematyzator (ang. *stemmer*) zawsze jako wynik dawał prawidłową formę podstawową wyrazu w rozumieniu lingwistycznym. Wystarczy, aby był to jeden, ten sam ciąg znaków. Omawiany etap ma szczególnie ważną rolę w przypadku dokumentów w języku polskim, który ma bardzo rozbudowaną fleksję. Jeśli nie zostanie wykorzystane sprowadzanie wyrazów do formy podstawowej, liczba cech znacznie wzrośnie.

4.2. Funkcje istotności atrybutów

Ograniczenie liczby atrybutów można także osiągnąć przez zastosowanie funkcji istotności, która stwierdza, czy dany atrybut jest ważny, czy można go pominąć.

Jedną z prostszych funkcji istotności atrybutu jest funkcja ważąca częstością termów (ang. *term frequency*) [17] – nazwana na potrzeby tej pracy TFIDF1.

$$\gamma(w_i, d_j) = tf_{ij} \tag{1}$$

Wartość $tf_{i,j}$ oznacza częstość wystąpień atrybutu w_i w dokumencie d_j, df_i to liczba dokumentów, w których występuje atrybut w_i. Natomiast cf_i określa liczbę wystąpień atrybutu w_i we wszystkich dokumentach. N określa liczbę wszystkich dokumentów w systemie, a M oznacza liczbę wszystkich atrybutów, jakie mogą potencjalnie wystąpić w dokumentach.

Problemem przy stosowaniu tej funkcji jest aprobowanie słów typu *stopword*, gdyż występują one bardzo często w dokumentach.

Wyrazy istotne dla analizy z reguły występują często w pojedynczych dokumentach, co wykorzystuje kolejna funkcja. Cała ich grupa nazywana jest funkcjami ważącymi częstością termów – odwrotną częstością w dokumentach (ang. *term frequency-inverse document frequency*) [14] – TFIDF2.

$$\gamma(w_i, d_j) = \frac{tf_{ij}}{df_i} \tag{2}$$

Lepsze efekty można osiągnąć, przez zastosowanie skalowania logarytmicznego – TFIDF3:

$$\gamma(w_i, d_j) = (1 + \log(tf_{ij}))\log\frac{N}{df_i} \tag{3}$$

W celu uwzględnienia długości dokumentu wprowadza się normalizację – TFIDF4:

$$\gamma(w_i, d_j) = \frac{\left(1 + \log(tf_{ij})\right)\log\frac{N}{df_i}}{\sqrt{\sum_{k=1}^{M}\left(\left(1 + \log(tf_{kj})\right)\log\frac{N}{df_k}\right)^2}} \tag{4}$$

Do wyznaczenia funkcji istotności można także wykorzystać entropię :

$$\gamma(w_i, d_j) = \left(\left(1 + \log(tf_{ij})\right)\cdot\left(1 + \frac{1}{\log N}\sum_{k=1}^{N}\left(\frac{tf_{ij}}{cf_i}\log\frac{tf_{ij}}{cf_i}\right)\right)\right) \tag{5}$$

5. Klasyfikacja dokumentów tekstowych w języku polskim

W klasyfikacji dokumentów w języku polskim występuje zjawisko „zanieczyszczenia" tekstu zwrotami z innych języków. Można wyróżnić trzy grupy tego zjawiska.

Pierwszy rodzaj obejmuje sentencje, zdania pochodzące z języka angielskiego. Mogą to być np. fragmenty komunikatów pewnego oprogramowania, a więc poprawne w sensie gramatycznym zdania. Do tej grupy można zaliczyć także cytowania wypowiedzi innych osób, które to, w przypadku cytowania dosłownego, również stanowią zazwyczaj odrębne frazy w języku obcym.

Cechą charakterystyczną tej grupy „zanieczyszczenia" jest możliwość wyodrębnienia części tekstu, który napisany jest w innym języku niż cała wiadomość. Kłopotliwa może być sytuacja, w której takie wtrącenie stanowi większość wiadomości. Wówczas pominięcie części obcojęzycznej w analizie może skutkować gorszą jakością klasyfikacji.

Drugi przypadek to wtrącanie angielskich zwrotów zarówno w sentencje w języku polskim, jak i jako oddzielne zazwyczaj równoważniki zdań.

Ta kategoria „zanieczyszczeń" powoduje mniejszą utratę wiadomości w przypadku pominięcia jej w analizie. Z drugiej strony, uwzględnienie tego rodzaju wtrąceń może pomóc w łatwiejszym wykryciu wiadomości pochodzących od konkretnej osoby, która używa tego rodzaju zwrotów. A wiadomości te mogą być dla użytkownika ważne. Dodatkowo, taki styl może charakteryzować większe grupy użytkowników, których informacje są uznawane za wartościowe. Przykładem mogą być programiści, którzy mogą umieszczać w swojej korespondencji wyrażenia anglojęzyczne ze względu na to, iż słowa kluczowe w językach programowania pochodzą z tego języka.

Trzeci rodzaj wtrąceń to używanie spolszczonych zwrotów angielskich. Tego rodzaju wtrącenia także stanowią wartościową informację dla systemu klasyfikującego, gdyż pozwalają na łatwiejsze rozpoznanie wiadomości pochodzących od pewnych grup użytkowników. W przypadku niewielkiej liczby wtrąceń spolszczonych zwrotów angielskich do wiadomości

pominięcie ich w analizie nie stanowi dużej utraty informacji. Ale można sobie wyobrazić użytkowników z grupy, która bardzo intensywnie posługuje się tego typu zwrotami. Wówczas pominięcie tej kategorii wyrażeń może spowodować znaczne zubożenie informacji. Skrajnym przypadkiem może być całkowita utrata zawartości wiadomości.

5.1. Proponowane podejście

Niniejsza sekcja przedstawia elementy, które należy wziąć pod uwagę przy klasyfikowaniu „zanieczyszczonych" tekstów. Część z nich to standardowe działania wykonywane podczas tworzenia reprezentacji dokumentów tekstowych, które przystosowane zostały jednak do omawianej sytuacji.

5.1.1. Usuwanie nieznaczących słów

System można wyposażyć w stop-listę dla każdego z języków – polskiego i angielskiego. Wówczas pozwoli to na usunięcie z fragmentów w języku angielskim słów występujących często i nieniosących ze sobą informacji, np. zaimek „the".

Należy w tej sytuacji zwrócić uwagę na występowanie słów o takiej samej pisowni w obu językach. Przykładem może być „i" oraz „I", które różnią się tylko wielkością litery. W języku polskim jest to spójnik, który zwyczajowo jest usuwany z dokumentów, a w angielskim to zaimek osobowy, który także trafia do stop-listy. Jest więc to przypadek, którego rozwiązanie nie stanowi problemu, gdyż może on znaleźć się na obu stop-listach.

W przypadku słów umieszczanych na stop-liście dla jednego języka nie jest to znaczący problem, gdyż powtarzać się mogą słowa krótkie, które z dużym prawdopodobieństwem znajdą się także na stop-liście drugiego języka, jak ma to miejsce w przypadku przedstawionego przykładu.

Dla wyrazów długich, które mogłyby nie znaleźć się w drugiej stop-liście, jest małe prawdopodobieństwo wystąpienia słów o jednakowej pisowni w dwóch językach.

5.1.2. Sprowadzanie słów do formy podstawowej

Uwzględnienie kwestii sprowadzania słów z języka angielskiego do formy podstawowej może pozytywnie wpłynąć na klasyfikację wiadomości elektronicznych w języku polskim, które zawierają wtrącenia w języku obcym.

Jednym z możliwych rozwiązań jest określenie, do jakiego języka należy dane słowo i użycie odpowiedniego lematyzatora. Wiąże się to jednak z koniecznością przydzielenia danego słowa do konkretnego języka, co może nie być łatwym zadaniem w przypadku krótkich wtrąceń.

Innym podejściem jest wykorzystanie lematyzatora dla języka polskiego (przy założeniu że większa część korespondencji pisana jest właśnie w tym języku) i wykorzystanie lematy-

zatora dla języka angielskiego dopiero wówczas, gdy pierwszy nie zmieni analizowanego wyrazu.

5.1.3. Przestrzeń atrybutów

Analizując wiadomości, w których mogą pojawić się fragmenty w innym języku, celowe jest określenie przestrzeni atrybutów, która obejmuje wyrazy (ich podstawowe formy) z obu języków. Wówczas możliwa będzie analiza całej zawartości wiadomości.

5.1.4. Wyrazy o jednakowej pisowni

W przypadku analizy tekstów w dwóch językach, polskim i angielskim, występują wyrazy o tej samej pisowni, ale różnym znaczeniu. Przykładem może być słowo „data". Przy zaproponowanym rozwiązaniu wystąpienia tego słowa w obu językach zostaną uznane za jeden atrybut.

Rozwiązaniem może być próba określenia, do którego języka należy dane wystąpienie słowa. Problemem przy takim podejściu może być trudność w określeniu przynależności do języka, gdy słowo jest krótkim wtrąceniem.

Próbę rozpoznawania języka danej wiadomości można podjąć w przypadku klasyfikacji dokumentów tekstowych w dwóch językach, gdzie wyrazy pochodzące z języka angielskiego będą występować częściej. Będzie to przedmiotem dalszych badań.

Sytuację tę można porównać do przypadków, w których występują słowa (w danym języku) o jednakowej pisowni i brzmieniu, ale różnych znaczeniach – homonimy. Na przykład „zamek" oznaczający budowlę będzie traktowany jako jeden atrybut razem z „zamkiem" błyskawicznym.

6. Wyniki eksperymentów

Do sprawdzenia praktycznej przydatności poszczególnych reprezentacji dokumentów tekstowych, w tym szczególnie do klasyfikacji poczty elektronicznej w języku polskim z wtrąceniami w języku angielskim, wykorzystany został naiwny klasyfikator bayesowski [11], który daje dobre wyniki w klasyfikacji dokumentów tekstowych. Zastosowane zostały proste reprezentacje dokumentów (częstościowa bez normalizacji oraz częstościowa z normalizacją) i funkcje istotności atrybutów, gdyż te bardziej skomplikowanie nie zawsze dają lepsze wyniki [22]. System wykorzystuje lematyzator (ang. *stemmer*) Portera [18] dla języka angielskiego oraz stworzony przez D. Weissa lematyzator dla języka polskiego [12] wykorzystujący automat skończony. System został zaimplementowany w języku Java.

Eksperymenty przeprowadzono na dwóch zbiorach danych tekstowych. Pierwszy to artykuły agencji Reuters (w języku angielskim), które wykorzystane zostały głównie do zweryfikowania poprawności implementacji klasyfikatora oraz do sprawdzenia skuteczności klasyfikacji przy wykorzystaniu naiwnego klasyfikatora bayesowskiego, lematyzatora Portera oraz rozważanych sposobach reprezentacji dokumentów i funkcji istotności atrybutów. Przeprowadzono przetwarzanie wstępne tego zbioru pozwalające na ograniczenie liczby kategorii. Drugi zbiór to autentyczne wiadomości otrzymane w wyniku korespondencji prowadzonej za pomocą poczty elektronicznej. Są to dokumenty tekstowe w języku polskim z wtrąceniami w języku angielskim. Miały one stworzyć warunki jak najbardziej zbliżone do praktycznych zastosowań. Obie kolekcje podzielone zostały na dwie części: zbiór trenujący (używany do uczenia) oraz zbiór testujący (wykorzystywany do sprawdzania poprawności klasyfikacji).

6.1. Klasyfikacja artykułów

Klasyfikator oparty na dyskretnych wartościach atrybutów (wykorzystujący reprezentację częstościową bez normalizacji) jest bardziej skuteczny przy klasyfikacji artykułów w języku angielskim od klasyfikatora wykorzystującego atrybuty znormalizowane. Klasyfikator „dyskretny" bez stosowania funkcji ograniczających reprezentację osiąga jakość (mierzoną współczynnikiem $F_{0,5}$)[1] ok. 0,9, natomiast klasyfikator „ciągły" ok. 0,7. Przy użyciu funkcji oceny istotności atrybutu widać także przewagę klasyfikatora wykorzystującego dyskretne wartości.

Zakładając, że odrzucamy jakość klasyfikacji poniżej progu 0,9, dla klasyfikatora „dyskretnego" najmniejszy rozmiar reprezentacji otrzymano dla: entropii (reprezentacja stanowiła około 26% rozmiaru zbioru wejściowego) oraz TFIDF4 (około 32% rozmiaru zbioru wejściowego). Przyjmując, że odrzucamy jakość klasyfikacji poniżej progu 0,7, dla klasyfikatora ciągłego najmniejszy rozmiar reprezentacji otrzymano dla TFIDF3 – około 25% rozmiaru zbioru wejściowego, a jakość klasyfikacji jest wyższa niż dla pełnego zbioru wejściowego.

6.2. Klasyfikacja poczty elektronicznej

Tabele 1 i 2 przedstawiają wyniki klasyfikacji poczty elektronicznej w języku polskim z wtrąceniami w języku angielskim – wykrywania niepożądanych wiadomości – dla różnych funkcji istotności atrybutów oraz progów odcięcia. Kursywą i pogrubieniem zaznaczono najlepsze wyniki dla współczynnika jakości F.

Tabela 1

Porównanie wyników klasyfikacji poczty elektronicznej (atrybuty dyskretne) dla wybranych funkcji istotności atrybutów i progów

[1] Definicję wszystkich miar użytych w eksperymentach można znaleźć w [2, 10, 19].

		Funkcja oceny istotności atrybutu			
		Brak (bez stopword)	Brak	TFIDF1	TFIDF4
				2	0,1
Rozmiar [KB]		598	542	215	66
Mikrouśrednianie	Precyzja	0,9764	0,9764	0,9647	0,9529
	Zupełność	0,9764	0,9764	0,9647	0,9529
	Dokładność	0,9764	0,9764	0,9647	0,9529
	Zaszumienie	0,0235	0,0235	0,0352	0,0470
	Jakość (F0,5)	***0,9764***	***0,9764***	0,9647	0,9529
Makrouśrednianie	Precyzja	0,9756	0,9756	0,9642	0,9600
	Zupełność	0,9756	0,9756	0,9642	0,9600
	Dokładność	0,9764	0,9764	0,9647	0,9529
	Zaszumienie	0,0217	0,0217	0,0326	0,0512
	Jakość (F0,5)	***0,9756***	***0,9756***	0,9642	0,9600

Przy klasyfikacji wykorzystane zostało podejście zaproponowane w niniejszej pracy. Wykorzystano więc dwie stop-listy – dla języka polskiego i angielskiego. Jako pierwszy wykorzystywany był lematyzator dla języka polskiego. Gdy nie zmieniał on analizowanego słowa, wówczas korzystano z lematyzatora dla języka angielskiego. Wykorzystano połączoną dla obu języków przestrzeń atrybutów bez rozdzielania termów pochodzących z różnych języków o tej samej pisowni i innym znaczeniu. Znaczniki HTML i SMTP nie były usuwane z wiadomości, co miało pomóc w rozpoznawaniu formatowania dokumentów.

Podobnie jak przy klasyfikacji artykułów klasyfikator oparty na dyskretnych wartościach atrybutów okazał się bardziej skuteczny od klasyfikatora wykorzystującego atrybuty znormalizowane (zachowano tu te same miary skuteczności w celu umożliwienia także porównania dwóch eksperymentów dla różnej tematyki zbiorów trenujących). Porównując uzyskane wyniki z pierwszym eksperymentem, można stwierdzić, że uzyskano zbliżone rezultaty dla klasyfikatora „dyskretnego" z niewielką przewagą jakości klasyfikowania poczty elektronicznej. Natomiast dla klasyfikatora „ciągłego" w jednym przypadku przy klasyfikacji wiadomości elektronicznych udało się uzyskać znacznie lepszy wynik niż przy analizowaniu artykułów. Okupione to było jednak znacznym rozmiarem zbioru trenującego.

Przy analizie poczty klasyfikator „dyskretny" nie wykazał zmian przy wykorzystaniu stop-listy (wyniki są identyczne). Natomiast ograniczenie reprezentacji za pomocą funkcji TFIDF1 z progiem 2 pozwoliło na zmniejszenie zbioru trenującego do objętości 215 kB (ok. 36% wyjściowego) przy zachowaniu jakości na poziomie 0,96. Dalsze ograniczenie zbioru trenującego do 66 kB za pomocą TFIDF4 (próg 0,1) zmniejszyło nieznacznie jakość do ok. 0,95 jednocześnie uzyskując 11% objętości zbioru wyjściowego.

Tabela 2

Porównanie wyników klasyfikacji poczty elektronicznej (atrybuty ciągłe) dla wybranych funkcji istotności atrybutów i progów

Funkcja oceny istotności atrybutu

		Brak (bez stopword)	Brak	TFIDF1	Entropia
				0,005	0,0005
Rozmiar [KB]		1930	1750	417	566
Mikrouśrednianie	Precyzja	0,4705	0,8941	0,4588	0,4588
	Zupełność	0,4705	0,8941	0,4588	0,4588
	Dokładność	0,4705	0,8941	0,4588	0,4588
	Zaszumienie	0,5294	0,1058	0,5411	0,5411
	Jakość (F0,5)	0,4705	***0,8941***	0,4588	0,4588
Makrouśrednianie	Precyzja	0,7321	0,9019	0,4588	0,4588
	Zupełność	0,7321	0,9019	0,4588	0,4588
	Dokładność	0,4705	0,8941	0,4588	0,4588
	Zaszumienie	0,4891	0,1114	0,5000	0,5000
	Jakość (F0,5)	0,7321	***0,9019***	0,4588	0,4588

Klasyfikator „ciągły” przy zastosowaniu redukcji atrybutów powodującej zmniejszenie objętości zbioru trenującego do ok. 30% osiągał jakość na poziomie 0,5. Podobny wynik uzyskany został dla pełnej reprezentacji (bez stop-listy). Zadziwiającą skutecznością okazał się klasyfikator przy wykorzystaniu stop-listy. Umożliwiło to uzyskanie jakości na poziomie 0,90.

Wnioskiem z przeprowadzonych eksperymentów może być spostrzeżenie, że sposób doboru zbioru trenującego (wybór funkcji ograniczającej i progu oraz użycia stop-listy) jest uzależniony od zastosowania klasyfikatora – tematyki, której ma sprostać klasyfikator.

7. Podsumowanie i dalsze prace

Przeprowadzone eksperymenty pokazują, że można w skuteczny sposób wykrywać niechciane wiadomości elektroniczne w języku polskim z angielskimi wtrąceniami.

Klasyfikator bayesowski, zastosowane proste reprezentacje dokumentów oraz funkcje istotności atrybutów pozwoliły na osiągnięcie wysokiej jakości klasyfikacji przy jednoczesnym ograniczeniu reprezentacji dokumentów.

Dalsze badania będą obejmowały sprawdzenie przedstawionego rozwiązania w klasyfikacji wiadomości poczty elektronicznej w języku polskim i angielskim jednocześnie oraz sprawdzenie przydatności innych klasyfikatorów w rozpoznawaniu niechcianej poczty.

BIBLIOGRAFIA

1. Barnbrook G.: Defining Language: A Local Grammar of Definition Sentences. John Benjamins 2002.

2. Bolc L., Cytowski J.: Modern Search Methods. Instytut Podstaw Informatyki PAN, Warszawa 1992.
3. Boratyński D.: Metody klasyfikacji dokumentów tekstowych w języku polskim. W: Wyzwania gospodarki elektronicznej – stan i perspektywy, Red. Tadeusz Grabiński, WSPiM, Chrzanów 2005.
4. Chakrabarti S.: Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann, San Francisco 2003.
5. Chrabąszcz M., Gołębski M., Bembenik R.: Metody klasyfikacji dokumentów tekstowych. Informatyka Teoretyczna i Stosowana, Wydawnictwo Politechniki Częstochowskiej, nr 3, Częstochowa 2002, s. 89÷100.
6. Church K. W., Gale W. A.: Inverse document frequency (IDF): A measure of deviations from Poisson. Proceedings of the Third Workshop on Very Large Corpora (WVLC), s. 121÷130, 1995.
7. Fang H., Tao T., Zhai C.: A formal study of information retrieval heuristics. Proceedings of SIGIR, 2004, s. 49÷56.
8. Fawcett T.: In vivo spam filtering: a challenge problem for KDD. SIGKDD Explorations 5(2) , 2003, s. 140÷148.
9. Feldman R., Sanger J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, New York 2007.
10. Gawrysiak P.: Automatyczna klasyfikacja dokumentów, Praca Doktorska, 2001.
11. Hastie T., Tibshirani R., Friedman J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York 2001.
12. http://www.cs.put.poznan.pl/dweiss/xml/projects/lametyzator/index.xml
13. Jain A., Murty M., Flynn P.: Data Clustering: A Review. ACM Computing Surveys, 31(3), wrzesień 1999.
14. Kroon de H., Mitchell T., Kerckhoffs E.: Improving learning accuracy in information filtering. International Conference on Machine Learning – Workshop on Machine Learning Meets HCI (ICML-96), 1996.
15. Liu R., Lu Y.: Incremental context mining for adaptive document classification. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002, s. 599÷604.
16. Mani I., Maybury M. T.: Advances in Automatic Text Summarization. MIT Press 2001.
17. Ponte J. M., Croft W. B.: A Language Modeling Approach to Information Retrieval. Research and Development in Information Retrieval, 1998, s. 275÷281.
18. Porter M. F.: An Algorithm for Suffix Stripping, Program, 14(3), 1980, s. 130÷137.
19. Rijsbergen Van C. J.: Information Retrieval. Dept. of Computer Science, University of Glasgow, 1979.

20. Robertson S. E.: Understanding inverse document frequency: On theoretical arguments for IDF. Journal of Documentation, 60(5), 2004, s. 503÷520.
21. Robertson S. E., Jones K. S.: Simple proven approaches to text retrieval. Tech. Rep. TR356, Cambridge University Computer Laboratory, 1997.
22. Salib M., Slicer M.: Spam Classification with Naive Bayes and Smart Heuristics, 2002.
23. Saul L., Pereira F.: Aggregate and Mixed-Order Markov Models for Statistical Language Processing, Association for Computational Linguistics, New Jersey 1997.
24. Scime A.: Web Mining: Applications and Techniques. Idea Group Inc (IGI) 2005.
25. Shakeri S., Rosso P.: Spam Detection and Email Classification. Information Assurance and Computer Security, IOS Press, 2006.
26. Song F., Croft W. B.: A General Language Model for Information Retrieval (poster abstract). Eighth International Conference on Information and Knowledge Management (CIKM'99), 1999.
27. Spärck Jones K.: IDF term weighting and IR research lessons. Journal of Documentation 60, 2004, s. 521÷523.
28. Stefanowski J., Zienkowicz M.: Classification of Polish Email Messages: Experiments with Various Data Representations. ISMIS 2006, s. 723÷728.
29. Willett P.: Recent Trends in Hierarchic Document Clustering: A Critical Review. Information Processing and Management, 24(5), 1988, s. 577÷597.
30. Youn S., McLeod D.: Efficient Spam Email Filtering using Adaptive Ontology. International Conference on Information Technology (ITNG'07), 2007, s. 249÷254.
31. Zorkadis V., Karras D. A., Panayotou M.: Efficient information theoretic strategies for classifier combination, feature extraction and performance evaluation in improving false positives and false negatives for spam e-mail filtering. Neural Networks 18(5-6), 2005, s. 799÷807.

Recenzent: Dr hab. inż. Andrzej Kwiecień, prof. Pol. Śląskiej

Wpłynęło do Redakcji 20 stycznia 2009 r.

Abstract

In classification of text documents several types of representations of documents are used. The aim is to build such a representation that is enough small and gives high accuracy of classification. A small representation means less data to store and process, i.e., less time

needed to process collected documents. Unfortunately, there is a trade-off between the size of representation and accuracy of classification. Too small representation gives poor results. Thus, a compromise between the size of a representation and the accuracy is needed.

This paper presents the representations of text documents (unigram, n-gram, γ-gram) and ways of their reduction in the case of SPAM detection in Polish with English phrases. An altered approach to building a representation of text documents is proposed. Several term weighting functions (including TF-IDF – expressions (1)-(4) and function based on enthropy – expression (5)) are used.

Proposed approach has been tested with Naïve Bayes classifier on a real e-mails data set. The results of the experiments (Tables 1 and 2) show that Naïve Bayes classifier and binary uniform representation give the high accuracy of classification of SPAM in Polish with English phrases.

Adres

Piotr ANDRUSZKIEWICZ: Politechnika Warszawska, Instytut Informatyki, ul. Nowowiejska 15/19, 00-665 Warszawa, Polska, P.Andruszkiewicz@ii.pw.edu.pl .