NAUKI PODSTAWOWE
CZASOPISMO TECHNICZNE
TECHNICAL TRANSACTIONS
FUNDAMENTAL SCIENCES
WYDAWNICTWO POLITECHNIKI KRAKOWSKIEJ
1-NP/2012
ZESZYT 18
ROK 109
ISSUE 18
YEAR 109

TOMASZ GĄCIARZ, KRZYSZTOF CZAJKOWSKI*

WPŁYW METODY SELEKCJI SŁÓW KLUCZOWYCH NA SKUTECZNOŚĆ KLASYFIKACJI STRON INTERNETOWYCH Z WYKORZYSTANIEM ALGORYTMU BOOSTINGU

THE INFLUENCE OF THE KEYWORDS SELECTION METHOD ON THE EFFECTIVENESS OF THE WEB PAGES CLASSIFICATION USING THE BOOSTING ALGORITHM

Streszczenie

Artykuł porusza zagadnienia dotyczące klasyfikacji stron internetowych. Klasyfikacja przeprowadzana jest w oparciu o analizę struktury oraz zawartości stron. Pod uwagę brane są cechy zróżnicowanym charakterze, w tym między innymi cechy strukturalne, wizualne, tekstowe, łączy internetowych. Przy budowie klasyfikatorów wykorzystano algorytm AdaBoost. Skupiono się na wpływie metody selekcji słów kluczowych na skuteczność procesu klasyfikacji.

Słowa kluczowe: strona internetowa, ekstrakcja cech, klasyfikacja, AdaBoost

Abstract

The paper concerns the issues of web pages analysis process. The classification is performed based on the analysis of the structure as well content of pages. Various characteristics are taken into account including inter alia, structural, visual, text, web and links features. During the construction of classifiers the AdaBoost algorithm was applied. This paper focuses on the impact of keyword selection methods on the effectiveness of the classification process.

Keywords: web page, features extraction, classification, AdaBoost

* Dr inż. Tomasz Gąciarz, mgr inż. Krzysztof Czajkowski, Instytut Teleinformatyki, Wydział Fizyki, Matematyki i Informatyki, Politechnika Krakowska.

1. Wstęp

Wraz ze wzrostem liczby stron internetowych coraz większym problemem staje się konieczność przeszukiwania dużych liczby stron internetowych. Dostępne wyszukiwarki internetowe pozwalające za pomocą podanych przez użytkownika słów kluczowych zaprezentować tylko te dokumenty, które spełniają zadane kryteria, nie są rozwiązaniem zadowalającym. Użytkownik może za ich pomocą określić na przykład, że interesują go informacje ze stron związanych z konkretną tematyką, ale tylko o określonym charakterze. Wydawać by się mogło, że rozwiązanie jest bardzo proste, wymaga tylko wpisania, poza konkretnym wyrażeniem, dodatkowego hasła (np. słowa „sklep"), określają charakter poszukiwanych witryn. Prostota tego zagadnienia (i rozwiązania) jest tylko pozorna, ponieważ słowa i całe wyrażenia pojawiają się na różnych stronach i nie koniecznie (lub nie całkowicie) muszą być związane z charakterem konkretnej strony.

Uzasadniona wydaje się więc próba skatalogowania różnych „rodzajów" (*genre*) stron i przypisania ich do właściwej im kategorii lub inaczej klasy przynależności. Strony należące do danej klasy charakteryzować się będą podobnym „stylem" jeśli chodzi o formę przekazu lub sposób prezentacji zawartości. Strony o podobnej treści będziemy mogli przypisać do różnych kategorii w sensie, w jakim je tu rozróżniamy. Wiele prac związanych z automatyczną klasyfikacją stron internetowych akcentuje tę ortogonalność treści i formy [10].

Z uwagi na liczbę stron oraz fakt, że ta liczba stale wzrasta konieczne jest opracowanie rozwiązań automatyzujących ten proces i umożliwiających cykliczne jego powtarzanie. W tym zakresie prowadzone są liczne prace obejmujące wykorzystanie różnych podejść sztucznej inteligencji, w tym między innymi zbiorów przybliżonych (*Rough Set*) [2, 3], uczenia maszynowego (*Machine Learning*) [5], algorytmów mrówkowych (*Ant Colony*) [6], naiwnych klasyfikatorów bayesowskich (*Naive Bayes*) [7], maszyn wektorów nośnych (*Support Vector Machine*) [8]. Zależnie od wielu czynników, w tym między innymi od przyjętej metody, rozważanej liczby klas (kategorii), wykorzystywanej liczby stron w zbiorze uczącym, uwzględniania języka stron, uzyskiwano różną skuteczność. Wciąż jednak nie opracowano rozwiązania, którego skuteczność byłaby satysfakcjonująca, a pracę nad różnymi podejściami nadal trwają.

W pracy [14] wykorzystano podejście opierające się na metodzie boostingu. Jest to metodą generowania zestawu komitetów klasyfikatorów. Charakteryzuje się wysokim (*state-of-the-art*) poziomem efektywności i solidnymi podstawami teoretycznymi z zakresu inteligentnych systemów uczących się. Jej skuteczności dowiedziono w rozwiązaniach szerokiego wachlarza problemów - m.in. automatycznej klasyfikacji tekstów [13]. Zainspirowani tym faktem autorzy postanowili sprawdzić jedną z odmian boostingu – algorytm AdaBoost w odniesieniu do zadania klasyfikacji umożliwiającej podział stron internetowych na poszczególne kategorie. Opracowane rozwiązanie bazuje na dużej liczbie różnorodnych cech opisujących dokumenty [14, 15]. Należy zwrócić uwagę na fakt, że wiele specyficznych algorytmów w języku angielskim nie sprawdza się w analizie języka polskiego.

W artykule skupiono się na weryfikacji wpływu wyboru metody selekcji słów kluczowych dla poszczególnych kategorii stron, na skuteczność procesu klasyfikacji. Metoda doboru słów kluczowych, będąca jednym z istotnych elementów opracowywanego systemu [15], wybrana zostanie w oparciu o przeprowadzone eksperymenty.

2. Kategorie stron internetowych

Skuteczność procesu klasyfikacji silnie zależy od wybranych klas, ich liczby oraz możliwie jak najbardziej niezależnych cech je charakteryzujących. Obecnie coraz trudniej jest wskazać zarówno takie kategorie, jak i cechy, ponieważ zawartość witryn internetowych jest często „wymieszana", dynamiczna i trudna do precyzyjnego określenia. Wyszukiwane cechy dotyczą zwykle języka i zawartości strony, formy oraz jej funkcjonalności.

W publikowanych pracach decydowano się na różne zestawy kategorii, kierując się różnymi kryteriami. W pracy [1] skupiono się na czterech klasach: FAQ, News, E-Shopping, Personal Home Pages. Wykorzystano 1280 przykładowych stron, po 170 stron dla każdej z czterech klas, oraz 600 stron nienależących do żadnej z rozpatrywanych klas. W artykule [10] zaproponowano podział na 8 gatunków: *link collection, help, shop, por-trayal non-private, portrayal private, article, download, discussion*. Wykorzystano 1209 stron internetowych podzielonych na 8 zbiorów (zgodnie z rozpatrywanymi klasami). Z każdego zbioru losowano po 100 stron i tylko one brały udział w poszczególnych ekspe-rymentach. W pracy [4] rozpatrywano 7 klas: *blog, eshop, FAQ, online newspaper front page, listing, personal home page, search page*. Wykorzystano zbiór 1400 stron internetowych, a każda klasa była reprezentowana przez 200 stron.

Niektóre cechy charakteryzują jednocześnie kilka klas, to znaczy ich obecność nie determinuje konkretnej klasy. Obecnie problem jest jeszcze bardziej złożony z uwagi na fakt, że strony internetowe stają się coraz bardziej rozbudowane, pełne elementów multimedialnych i są tworzone w coraz bardziej zaawansowanych technologiach. Nawet wówczas gdy, jak w przypadku stron typu FAQ, wciąż zachowana jest pewnego rodzaju „prostota" takich dokumentów, są one często częścią większych stron (forum, portali itp.), Sytuacja komplikuje się dodatkowo, gdy wybrana zostanie większa liczba klas, na jakie dzielone są strony. Zwiększając liczbę klas, coraz trudniej jest jasno i precyzyjnie wskazać zestaw kilku czy kilkunastu cech, jakie wyróżniają daną klasę na tle innych.

Internet cechuje nie tylko stały wzrost liczby stron (różnych klas – przy czym liczba stron poszczególnych klas wzrasta nierównomiernie), ale także ewolucja istniejących klas oraz pojawianie się klas zupełnie nowych [9].

Problem ten to zapewne podstawowa przyczyna, dla której w niektórych pracach (np. [1]) skupiono się na stosunkowo niewielkiej grupie kategorii. Pozwala to zazwyczaj na uzyskanie dobrych wyników pod kątem skuteczności. Pamiętając jednak o tym, że poza skutecznością drugim ważnym wyznacznikiem jest użyteczność, zawężanie się do kilku klas może okazać się niesatysfakcjonujące. Z uwagi na ten problem, w pracach wykorzystujących podział na większą liczbę klas (m.in. [4, 10]) wykorzystywano znacznie większą liczbę cech, w tym m.in. znaki interpunkcyjne, charakterystykę długości strony, różne tagi HTML itp. Prowadzono także eksperymenty na różnie skonstruowanych podgrupach cech.

W prezentowanych badaniach rozważano 9 klas. W eksperymentach wykorzystano w celach treningowych 1800 stron, po 200 dla każdej z 9 klas:

- Artykuł (*Article*) – wypowiedź publicystyczna;
- Blog (*Blog*) – zbiór odrębnych, samodzielnych, uporządkowanych chronologicznie wpisów, których twórcą jest właściciel strony;
- E-sklep (*E-shop*) – sklep internetowy;
- FAQ (*Frequently Asked Questions*) – zbiory „często zadawanych pytań"i odpowiedzi;
- Forum (*Forum*) – forma dyskusji, posiadająca wyodrębnione wątki;

– Katalog (*Catalog*)– moderowany ręcznie zbiór adresów stron internetowych, pogrupowany tematycznie;
– Portal (*Portal*) – serwis informacyjny dostępny z jednego adresu internetowego, rozbudowany o różnorodne funkcje internetowe;
– Strona domowa (*Personal Home Page*) – prywatna strona internetowa stanowiąca internetową wizytówkę danego użytkownika (właściciela);
– Strona firmowa (*Company Home Page*) – strona internetowa stanowiąca internetową wizytówkę danej firmy (będącej jej właścicielem);

3. Cechy opisujące strony internetowe

Skuteczna klasyfikacja stron internetowych opiera się na znalezieniu odpowiednich cech je charakteryzujących. Trudno jest określić z góry, które cechy są na pewno istotne (i okażą się kluczowe w procesie klasyfikacji), a które mają znaczenie marginalne. Wydaje się, że jedyną drogą weryfikacji, które atrybuty stron i w jakim stopniu są znaczące, są praktyczne testy. W omawianym podejściu przyjęto założenie, że wydobywana będzie możliwie duża liczba właściwości opisujących strony. W przypadku stron internetowych istotne cechy dotyczą zarówno treści stron (elementów widocznych dla odwiedzającego stronę), ich struktury (rodzajów i treści tagów html) oraz funkcjonalności (m.in. skrypty, linki do innych stron) [1].

Bardziej precyzyjnie, cechy opisujące stronę HTML można podzielić na kilka kategorii:

- Cechy tekstowe: statystyki słów kluczowych (zawartych w słownikach zbudowanych dla każdej kategorii), inne statystyki oparte o słowniki, ogólne statystyki tekstu, znaki interpunkcyjne, znaki typograficzne, statystyki części mowy. W prezentowanym rozwiązaniu skupiono się na słowach w języku polskim. Między innymi wybrano następujące cechy:
  – stosunek liczby wystąpień słowa kluczowego do wszystkich słów,
  – stosunek liczby wystąpień słów będących daną częścią mowy do wszystkich słów,
  – stosunek liczby wystąpień w tekście znaku interpunkcyjnego do wszystkich znaków interpunkcyjnych,
  – stosunek liczby wystąpień w tekście znaku typograficznego do wszystkich znaków typograficznych,
  – stosunek liczby wystąpień emotikony do wszystkich emotikon,
  – stosunek liczby wystąpień emotikon do wszystkich słów.
- Cechy strukturalne, m.in:
  – stosunek liczby tagów html do ogólnej liczby treści na stronie,
  – stosunek liczby wystąpień sekwencji tagów (tzw. *N*-gramów) do wszystkich tagów,
  – stosunek liczby kodu skryptowego do pozostałej treści,
  – stosunek liczby kodu skryptowego do liczby kodu html,
  – średnia liczba wystąpień poszczególnych tagów związanych ze strukturą dokumentu w odniesieniu do wszystkich tagów.
  – stosunek liczby wystąpień tagu (np. <td>) do wszystkich Tagów,
  – stosunek liczby wystąpień sekwencji tagów (np. <td><img>) do wszystkich tagów,
  – stosunek liczby słów do ilości treści,
  – stosunek liczby kodu css do ilości treści,
  – stosunek liczby wystąpień atrybutu (np. id) do wszystkich tagów,

  - wariancja wartości określonego atrybutu dla tagu (np. <script type=..>).
- Cechy wizualne, m.in.:
  - związane z formatowaniem – średnie liczby poszczególnych tagów formatujących,
  - związane z obrazami – stosunek liczby tagu <img> do wszystkich tagów, stosunki wystąpień obrazów w poszczególnych, typowych formatach, stosunki wystąpień obrazów o wielkościach: małych, średnich i dużych,
  - związane z plikami multimedialnymi – stosunki liczby plików w różnych formatach do liczby wszystkich plików multimedialnych,
  - związane ze stylem – w tym również występowanie odwołań do zewnętrznych arkuszy CSS.
- Cechy linków do innych stron, m.in.:
  - liczba wszystkich linków,
  - stosunek linków prowadzących do tej samej domeny do wszystkich linków,
  - stosunek linków prowadzących do innej domeny do wszystkich linków,
  - stosunek linków „mailowych" do wszystkich linków,
  - stosunek linków „obrazkowych" do wszystkich linków,
  - stosunek linków związanych z obrazami do wszystkich linków.

4. Metody pozyskiwania słów kluczowych

Jak przedstawiono w punkcie 3, jednym z najistotniejszych źródeł informacji o przynależności danej strony do konkretnej kategorii mogą być cechy tekstowe. Wśród nich szczególną rolę pełnią słowa kluczowe. Jest to spowodowane po pierwsze stosunkowo dużą liczbą otrzymanych w ten sposób cech (a więc i słabych klasyfikatorów) – aż 50 dla każdej z 9 rozpatrywanych klas. Po drugie są to jedyne cechy (sposób rozważanych) dotyczące treści stron, a nie ich struktury lub formatowania. Dodatkowo jest to jedyny podzbiór rozpatrywanych cech, którego liczebność można łatwo zmieniać przez ustawienie limitu rozpatrywanych słów.

W omawianym rozwiązaniu wzięto pod uwagę trzy różne sposoby pozyskiwania słów kluczowych. Każdą z trzech metod zaimplementowano w aplikacji i przeprowadzono szereg testów omówionych w punkcie 6. Ponadto aplikacja umożliwia zmianę liczby wyszukiwanych słów kluczowych, jednak tutaj przyjęto ich liczba na 50 dla każdej klasy i była to liczba stała podczas weryfikacji wszystkich trzech metod.

Metoda I (którą można określić jako „podejście ogólne") polega na ustaleniu liczby wystąpień danego słowa we wszystkich dokumentach znajdujących się w konkretnym zbiorze uczącym. Zbiorem uczącym jest w tym przypadku zbiór stron należących do jednej kategorii. Można powiedzieć, że zbiór wszystkich stron z danej kategorii traktowany jest jak jeden wielki dokument, a słowa kluczowe w nim zawarte są po prostu zliczane. Zaletą tej metody jest jej prostota, wadą natomiast może być fakt, że w przypadku pliku html (strony) zawierającego bardzo dużą liczbą sztucznie powtarzających się tych samych słów słowo takie może uzyskać wysoką pozycję w hierarchii, mimo że nie pojawia się w pozostałych plikach (stronach).

Metoda II (która może być określona jako metoda „weryfikacji obecności") opiera się na ustaleniu obecności danego słowa w konkretnym dokumencie (stronie WWW). Inaczej mówiąc, stwierdzane jest istnienie danego słowa na zasadzie „tak/nie", bez rozpatrywania, ile razy to słowo występuje. Metoda wydaje się nie być czuła na sztuczne nagromadzenie jednego wyrazu

(np. w celu pozycjonowania strony WWW) w jakimś dokumencie. Jednakże, nie biorąc pod uwagę częstotliwości występowania poszczególnych słów w dokumentach, celowo pozbawiamy się pewnej informacji. To, co stanowić może zaletę tej metody, może być również jej wadą.

Metoda III (którą można określić mianem „hybrydowej") stanowić ma z założenia złączenie obu powyższych metod (I i II), sumując ich zalety. W metodzie tej brana jest pod częstość występowania danego słowa w określonym dokumencie, lecz nie jako bezwzględne wystąpienia, ale procentowo, w stosunku do innych słów kluczowych z tego dokumentu. Z jednej więc strony nie ma utraty informacji o nasileniu występowania danego słowa (co może wynikać z mocnego skoncentrowania danej strony na jednym zagadnieniu, obiekcie, produkcie, itp.). Z drugiej strony liczba powtórzeń danego słowa nie jest brana pod uwagę wprost, ale jako procentowy udział w stosunku do innych słów z dokumentu, eliminowany jest w ten sposób efekt sztucznego nasycania konkretnej strony danym słowem.

Każda metoda daje w praktyce nieco różniące się od siebie wyniki. Zostały one zaprezentowane w rozdziale 6. Dzięki wykonanym eksperymentom zweryfikowano poszczególne metody, co pozwoliło na uzasadnienie wyboru jednej z nich do prac nad innymi aspektami opracowywanego systemu.

5. Aplikacja

5.1. Przygotowanie słowników kategorii

Na tym etapie analizy stron generowane są (charakterystyczne dla danej klasy decyzyjnej) słowniki zawierające słowa kluczowe. Słowa kluczowe dołączają następnie do ustalonego zbioru cech stron internetowych. Wydobycie słów kluczowych z dokumentu HTML jest zadaniem złożonym. W trakcie przetwarzania wstępnego dokumentu usuwane są zbędne znaczniki HTML, atrybuty HTML oraz wszystkie znaki niebędące słowami. Usuwa się także słowa, które zazwyczaj nie wnoszą żadnych informacji do tekstu, służą tylko łączeniu kolejnych treści (tzw. stop words). Listy takich słów dla języka angielskiego są ogólnie dostępne w Internecie. Dla języka polskiego konieczne jest utworzenie takiej listy samodzielnie. Następnie wszystkie słowa dostępne w dokumencie sprowadzane są do rdzenia słowotwórczego. Pozwala to rozpoznać występowanie danego słowa w tym samym dokumencie, ale w innej formie gramatycznej. Proces ten zwany *stemmingiem* lub lematyzacją jest stosukowo nieskomplikowany dla języka angielskiego, jest jednak dość złożony w przypadku języka polskiego (ze względu na jego skomplikowaną składnię, fleksję oraz ortografię). W aplikacji skorzystano z projektu „morfologik", który zawiera w sobie *stemmer* dla języka polskiego [11].

Na podstawie stałego zestawu cech strukturalnych, wizualnych, łączy oraz atrybutów tekstowych (rozszerzonych o słowniki słów kluczowych) tworzony jest wektor dla każdej strony internetowej. Wektory są normalizowane do przedziału (0,1).

5.2. Proces uczenia funkcji klasyfikującej

Dysponując zbiorem próbek stron WWW oraz ich reprezentacją w postaci wektora cech, przystępujemy do budowy klasyfikatorów przy pomocy algorytmu AdaBoost. Dla każdej zdefiniowanej kategorii konstruowany jest jeden tzw. „silny klasyfikator" będący kombina-

cją liniową „słabych klasyfikatorów" (najczęściej pojedynczych cech). Jego zadaniem będzie udzielenie odpowiedzi: czy i w jakim stopniu badana próbka testowa należeć będzie do tej kategorii, czy też bliżej jej będzie do całej reszty traktowanej jako inna kategoria. Będziemy tu mieć więc do czynienia z problemem decyzji o przynależności do jednej z dwóch klas. Nazwa „słaby klasyfikator" nawiązuje do faktu, że wymagamy od niego skuteczności tylko nieco lepszej niż losowa (>50%). W tym kontekście cecha, która pozwala nam z prawdopodobieństwem lepszym niż 50% wnioskować o przynależności strony do danej kategorii, spełnia wymagania słabego klasyfikatora. Silny klasyfikator związany z daną kategorią będzie dawał odpowiedź, czy dana próbka będzie należała do tej kategorii, czy też bliżej jej do całej reszty.

Wykorzystywany algorytm opublikowany został w 1995 roku (Y. Freund, R. Schapire – [12]). Autorzy udowodnili, że błąd silnego klasyfikatora końcowego maleje wykładniczo w kierunku zera. Jest to algorytm iteracyjnouczący, który w kolejnych krokach wybiera najlepsze „słabe" klasyfikatory, opierając się na zbiorze uczącym i dostępnych „słabych" klasyfikatorach. W każdym kolejnym kroku t słabe klasyfikatory ht są dobierane tak, żeby najbardziej skupiały się na przypadkach złego sklasyfikowania (algorytm po każdej rundzie zwiększa wagi źle sklasyfikowanych danych). Dodatkowo każdemu wybieranemu klasyfikatorowi przypisywana jest waga określająca jego ważność. Po zakończeniu działania algorytmu (po T krokach) otrzymujemy klasyfikator końcowy HT, którego obliczamy korzystając ze wzoru:

$$H_T(x) = \begin{cases} 1 \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 \text{ w przeciwnym razie} \end{cases}$$

Pseudokod algorytmu

1. Mając zbiór próbek stron (x_i, y_i), …,(x_N,y_N), gdzie y_i= 0,1 odpowiednio dla przykładów negatywnych (strony należące do wszystkich oprócz rozpatrywanej kategorii) i pozytywnych (strony należące do danej kategorii), każdemu elementowi przypisz wagę

$$d_i^{(1)} = \frac{1}{N}, i = 1,...,N$$

2. Dla kroków $t = 1,\ldots,T$,
   1) Wybierz klasyfikator $h_t : X \to \{0, +1\}$ minimalizujący błąd

$$\varepsilon_t = \sum_{n=1}^{N} d_i^{(t)} \left[y_n \neq h_t(x_n) \right]$$

   2) Oblicz $\alpha_t = \frac{1}{2} \ln \frac{1-\varepsilon_t}{\varepsilon_t}$

   3) Popraw wagi $d_i^{(t+1)} = \frac{d_i^{(t)} \exp\left\{-\alpha_t y_i h_t(x_i)\right\}}{Z_t}$, gdzie Z_t jest stałą normalizującą, taką, że $\sum_{i=1}^{N} d_i^{(t+1)} = 1$.

   4) Przerwij, jeśli $\varepsilon_t = 0$ lub $\varepsilon_t \geq 0{,}5$ i $T = t - 1$, jeśli nie, wróć do kroku 2

   5) Klasyfikator końcowy $H_T(x) = \begin{cases} 1 \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 \text{ w przeciwnym razie} \end{cases}$

6. Wyniki eksperymentów

W rozdziale zamieszczono wyniki eksperymentów polegających na wytrenowaniu klasyfikatorów z wykorzystaniem 200 stron dla każdej z 9 klas (w sumie 1800 stron), a następnie przetestowaniu skuteczności klasyfikacji za pomocą 30 stron dla każdej kategorii (270 stron) nie wykorzystywanych w procesie uczenia.

Rozpatrywano trzy warianty trenowania klasyfikatora, zależnie od zbioru cech, na podstawie których odbywał się ten proces. W każdym przypadku brany pod uwagę był ten sam, ogólny zestaw właściwości, składający się z cechy tekstowych, strukturalnych, wizualnych oraz linków (łączy). Różnica polegała na wyborze metody zbierania słów kluczowych, w oparciu o które konstruowany był zestawy cech tekstowych.

W pierwszym z rozpatrywanych przypadków wykorzystano metodę I (ogólną). Wyniki zaprezentowano w tabeli 1.

Tabela 1

Tabela krzyżowa skuteczności klasyfikacji stron testowych dla metody I

Rozpoznanie / Kategorie	Artykuł	Blog	E-sklep	FAQ	Forum internetowe	Katalog	Portal	Strona domowa	Strona firmowa	Niesklasyfik.
Artykuł	70%	3,33%	0%	0%	3,33%	0%	13,33%	0%	6,67%	3,33%
Blog	3,33%	76,67%	0%	0%	0%	0%	10%	3,33%	6,67%	0%
E-sklep	0%	0%	63,33%	6,67%	0%	3,33%	6,67%	6,67%	6,67%	6,67%
FAQ	13,33%	0%	6,67%	30%	3,33%	0%	16,67%	3,33%	20%	6,67%
Forum Internet.	0%	3,33%	6,67%	33,33%	46,67%	0%)	3,33%	3,33%	3,33%	0%
Katalog	13,33%	6,67%	16,67%	0%	0%	50%	10%	3,33%	0%	0%
Portal	20%	10%	3,33%	0%	0%	3,33%	43,33%	0%	20%	0%
Strona domowa	0%	3,33%	6,67%	3,33%	0%	0%	3,33%	26,67%	26,67%	30%
Strona firmowa	3,33%	10%	0%	3,33%	0%	6,67%	6,67%	26,67%	30%	13,33%

W tabeli 2 przedstawiono wyniki drugiego eksperymentu (również dla metody I). Obrazuje ona prawdopodobieństwa dobrego sklasyfikowania próbki, jeśli weźmie się pod uwagę jej wystąpienie w pierwszych dwóch lub pierwszych trzech najlepszych propozycjach zwróconych przez klasyfikatory. Miarą przynależności do danej klasy jest tutaj różnica

$\sum_{t=1}^{T} \alpha_t h_t(x) - \frac{1}{2}\sum_{t=1}^{T} \alpha_t$, (pod warunkiem, że jest ona nieujemna).

Jak można zauważyć, skuteczność rozpoznawania wszystkich klas jest znacząco lepsza. Jednak poprawa prawidłowości klasyfikacji nie jest jednakowa. Największą poprawę zaobserwowano dla kategorii: E-sklep (o 20%), Forum internetowe (20%), Portal (37%). Wyniki dla kategorii: FAQ, Strona domowa i Strona firmowa są wciąż najsłabsze.

Tabela 2

Skuteczność klasyfikacji stron testowych dla metody I

Kategorie / Skuteczność	Artykuł	Blog	E-sklep	FAQ	Forum internetowe	Katalog	Portal	Strona domowa	Strona firmowa	Średnia skuteczność
Pierwsze dwie propozycje	80%	83,33%	83,33%	36,67%	66,67%	63,33%	80%	33,33%	43,33%	63%
Pierwsze trzy propozycje	90%	83,33%	83,33%	46,67%	70%	66,67%	93,33%	33,33%	43,33%	68%

Wyniki wskazują na poprawę rozpoznawania większości kategorii (w granicach 10-13%), jednak w przypadku kategorii: Blog, E-sklep, Strona domowa, Strona firmowa, nie nastąpiła poprawa.

W kolejnym eksperymencie wykorzystano kolejną metodę selekcji słów kluczowych – metodę II. Wyniki zaprezentowano w tabeli 3.

Tabela 3

Tabela krzyżowa skuteczności klasyfikacji stron testowych dla metody II

Rozpoznanie / Kategorie	Artykuł	Blog	E-sklep	FAQ	Forum internetowe	Katalog	Portal	Strona domowa	Strona firmowa	Niesklasyfik.
Artykuł	56,67%	13,33%	0%	3,33%	0%	0%	20%	0%	6,67%	0%
Blog	6,67%	80%	3,33%	0%	0%	0%	3,33%	0%	6,67%	0%
E-sklep	0%	0%	50%	0%	3,33%	6,67%	6,67%	6,67%	16,67%	10%
FAQ	3,33%	0%	0%	43,33%	6,67%	3,33%	13,33%	6,67%	16,67%	6,67%
Forum Internet.	3,33%	0%	3,33%	13,33%	56,67%	0%	13,33%	3,33%	6,67%	0%
Katalog	0%	6,67%	16,67%	0%	0%	56,67%	13,33%	3,33%	0%	3,33%
Portal	3,33%	13,33%	0%	0%	0%	3,33%	63,33%	0%	16,67%	0%
Strona domowa	0%	0%	0%	0%	0%	0%	0%	30%	30%	40%
Strona firmowa	0%	3,33%	3,33%	0%	0%	6,67%	6,67%	26,67%	36,67%	16,67%

Analogicznie do pierwszego eksperymentu również w tym przypadku sprawdzono prawdopodobieństwo dobrego sklasyfikowania próbki, biorąc pod uwagę jej wystąpienie w pierwszych dwóch lub pierwszych trzech najlepszych propozycjach zwróconych przez klasyfikatory. Wyniki zamieszczono w tabeli 4.

Tabela 4

Skuteczność klasyfikacji stron testowych dla metody II

Kategorie / Skuteczność	Artykuł	Blog	E-sklep	FAQ	Forum internetowe	Katalog	Portal	Strona domowa	Strona firmowa	Średnia skuteczność
Pierwsze dwie propozycje	80%	80%	73,33%	60%	70%	83,33%	86,67%	33,33%	56,67%	69%
Pierwsze trzy propozycje	96,67%	83,33%	76,67%	63,33%	76,67%	83,33%	93,33%	33,33%	60%	74,07%

W trzecim eksperymencie, do trenowania klasyfikatora, wykorzystano cechy, wśród których znalazły się słowa kluczowe uzyskane metodą III („hybrydową"). Wyniki zamieszczono w tabeli 5.

Tabela 5

Tabela krzyżowa skuteczności klasyfikacji stron testowych dla metody III

Rozpoznanie / Kategorie	Artykuł	Blog	E-sklep	FAQ	Forum internetowe	Katalog	Portal	Strona domowa	Strona firmowa	Niesklasyfik.
Artykuł	50%	13,33%	0%	3,33%	0%	0%	26,67%	0%	6,67%	0%
Blog	10%	76,67%	0%	0%	0%	0%	10%	0%	3,33%	0%
E-sklep	3,33%	0%	56,67%	0%	0%	10%	0%	6,67%	13,33%	10%
FAQ	0%	3,33%	3,33%	43,33%	6,67%	0%	10%	6,67%	16,67%	10%
Forum Internet.	6,67%	0%	0%	6,67%	63,33%	0%	13,33%	3,33%	6,67%	0%
Katalog	0%	6,67%	13,33%	0%	0%	53,33%	16,67%	6,67%	0%	3,33%
Portal	10%	13,33%	0%	0%	0%	0%	50%	3,33%	23,33%	0%
Strona domowa	0%	0%	0%	0%	0%	0%	0%	30%	30%	40%
Strona firmowa	0%	3,33%	3,33%	0%	0%	0%	6,67%	26,67%	36,67%	23,33%

Tabela 6 prezentuje prawdopodobieństwa dobrego sklasyfikowania próbki, biorąc pod uwagę jej wystąpienie w pierwszych dwóch lub pierwszych trzech najlepszych propozycjach zwróconych przez klasyfikatory.

Tabela 6

Skuteczność klasyfikacji stron testowych dla metody IIII

Kategorie / Skuteczność	Artykuł	Blog	E-sklep	FAQ	Forum internetowe	Katalog	Portal	Strona domowa	Strona firmowa	Średnia skuteczność
Pierwsze dwie propozycje	83,33%	83,33%	70%	53,33%	66,67%	76,67%	83,33%	33,33%	53,33%	67,04%
Pierwsze trzy propozycje	100%	86,67%	73,33%	60%	76,67%	76,67%	93,33%	33,33%	53,33%	73 %

Porównanie średniej skuteczności dla pierwszej, pierwszych dwóch oraz pierwszych trzech propozycji, dla poszczególnych metod zbierania słów kluczowych, przedstawione zostało w tabeli 7.

Tabela 7

Skuteczność klasyfikacji stron testowych dla różnych metod

Metody / Średnia skuteczność	Metoda I	Metoda II	Metoda III
Pierwsza propozycja	49%	52,59%	51%
Pierwsze dwie propozycje	63%	69%	67,04%
Pierwsze trzy propozycje	68%	74,07%	73 %

Jak można zauważyć, średnia skuteczność klasyfikacji jest w przypadku metody II znacząco wyższa, niż w przypadku metody I (ogólnej) oraz nieznacznie wyższa w porównaniu z metodą III („hybrydową"). Różnica występuje dla pierwszej, pierwszych dwóch oraz pierwszych trzech propozycji.

7. Wnioski

W artykule omówiono zastosowanie metody boostingu w klasyfikacji stron internetowych. Skupiono się na wpływie rodzaju metody doboru słów kluczowych opisujących dokumenty, na skuteczność procesu klasyfikacji. Jest to jeden z czynników, który obok

wyboru klas, na jakie dzielone są strony, różnorodności cech opisujących strony, doboru stron do zbioru uczącego oraz rozmiaru tego zbioru, może rzutować na poprawność uzyskiwanych rezultatów.

Jak wynika z przeprowadzonych eksperymentów, wybór metody selekcji słów kluczowych podczas tworzenie zestawu cech tekstowych stanowiących dane wejściowe dla procesu trenowania klasyfikatora, ma istotne znaczenie. Pomimo iż cechy tekstowe stanowią tylko jedną z czterech grup rozpatrywanych cech, zmiana metody doboru słów kluczowych może skutkować zmianą skuteczności na poziomie od 2 do 7 %. W przypadku zmiany metody I (ogólnej) na metodę II, dla pierwszych trzech propozycji, poprawa wynosi ponad 7%, co stanowić zmianę poprawności wyników z 68% na 74,04%, a więc o ponad 1/10.

Kolejne eksperymenty skupiać się będą na dalszej poprawie skuteczności poprzez zastosowanie różnych metod zbierania słów kluczowych dla różnych kategorii. Można bowiem zaobserwować, że choć średnia skuteczność przemawia na korzyść metody II, to różne metody charakteryzują się różną skutecznością zależnie od rozpatrywanej kategorii strony (przed uśrednieniem) i decyzja o wyborze metody nie musi być tak oczywista. Chociaż więc najwyższą średnią skuteczność wykazuje metoda dokumentowa, to połączenie różnych metod (dla odpowiednich klas) może dać jeszcze lepsze rezultaty.

Literatura

[1] Dong L., Watters C., Duffy J., Shepherd M., *An Examination of Genre Attributes for Web Page Classification. Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, HICSS 2008.

[2] Yin S., Wang F., Xie Z., Qiu Y., *Study on Web-Page Classification Algorithm Based on Rough Set Theory*, Proceedings of ISIP'2008, 202-206.

[3] Czajkowski K., *Reguły decyzyjne i bazy danych w klasyfikacji stron internetowych*, Studia Informatica, Gliwice, Vol. 30, No. 2A(83), 2009, 355-372.

[4] Santi M., *Some issues in automatic genre classification of web pages*, Proceedings of JADT 2006.

[5] Tsukada M., Washio T., Metoda H., *Automatic Web-Page Classification by Using Machine Learning Methods*, Web Intelligence: Research and Development, LNAI 2001, Springer-Verlag, 303-313.

[6] Holden N., Freitas A.A., *Web Page Classification with an Ant Colony Algorithm*, Parallel problem solving from nature - PPSN VIII, LNCS 3242, Springer-Verlag 2004, 1092-1102.

[7] Fernandez V. F., Unanue R.M., Herranz S.M., Rubio A.C., *Naive Bayes Web Page Classification with HTML Mark-Up Enrichment*, International Multi-Conference on Computing in the Global Information Technology, 2006. ICCGI '06.

[8] Xue W., Huang W., Lu Y., *Application of SVM in Web Page Categorization*, IEEE International Conference on Granular Computing, 2006, 469-472.

[9] Shepherd M., Watters C., *Identifying Web Genre: Hitting A Moving Target*, Proc. of the WWW2004 Conference. Workshop on Measureing Web Searach Effectiveness: The User Perspective, New York, 18 May 2004.

[10] Meyer zu Eissen S., Stein B., *Genre Classification of Web Pages: User Study and Feasibility Analysis*, In: Biundo S., Fruhwirth T., Palm G. (Eds.): Advances In Artificial Intelligence, Springer 2004, 256-269.
[11] *Strona projektu Morfologik* (http://morfologik.blogspot.com).
[12] Freund Y., Schapire R.E., *A decision-theoretic generalization of on-line learning and an application to boosting*, In Computational Learning Theory: Eurocolt '95, Springer-Verlag, 1995, 23-37.
[13] Sebastiani F., Sperduti A., Valdambrini N., *An improved boosting algorithm and its application to automated text categorization*, Centre National de la Recherche Scientifique, 2000.
[14] Gąciarz T., Czajkowski K., Niebylski M., Szawernoga R., *Klasyfikacja stron internetowych z wykorzystaniem algorytmu boostingu*, Studia Informatica Vol. 32, No. 2A (96), 2011.
[15] Gąciarz T., Czajkowski K., Niebylski M., *Adaboost ranking results improvement by pairwise classifiers for web page classification*, Czachórski T., Kozielski S., Stańczyk U. (Eds.), [in:] Advances in Intelligent and Soft Computing, vol. 103, Man-Machine Interactions 2, Springer-Verlag Berlin Heidelberg, 2011.