STUDIA INFORMATICA 2011
Volume 32 Number 2A (96)

Tomasz GĄCIARZ, Krzysztof CZAJKOWSKI, Maciej NIEBYLSKI,
Ryszard SZAWERNOGA
Politechnika Krakowska, Wydział Fizyki, Matematyki i Informatyki,
Instytut Teleinformatyki,

KLASYFIKACJA STRON INTERNETOWYCH Z WYKORZYSTANIEM ALGORYTMU BOOSTINGU

Streszczenie. Artykuł dotyczy analizy informacji opisujących strony internetowe. Celem analizy jest wsparcie procesu ich klasyfikacji. Brane są pod uwagę cechy o zróżnicowanym charakterze, w tym między innymi cechy: strukturalne, wizualne, tekstowe, łączy internetowych. Przy budowie klasyfikatorów wykorzystano algorytm AdaBoost. W artykule przedstawiono implementację omawianego rozwiązania oraz wyniki przeprowadzonych eksperymentów.

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

WEB PAGES CLASSIFICATION USING BOOSTING ALGORITHM

Summary. The article concerns the analysis of information describing the web pages. The aim of the analysis is to support the process of their classification. 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. The paper presents the implementation of this solution and the results of experiments.

Keywords: web page, features extraction, classification, AdaBoost.

1. Wstęp

W sieci Internet dostępna jest ogromna liczba stron WWW, która nieustannie wzrasta, a tempo tego wzrostu wciąż się zwiększa. Jednym z podstawowych problemów, przed jakim staje użytkownik korzystający z sieci jest konieczność przeszukiwania wielu stron internetowych w poszukiwaniu interesujących go treści. Są oczywiście dostępne wyszukiwarki inter-

netowe, pozwalające, za pomocą podanych przez użytkownika słów kluczowych, zaprezentować tylko te dokumenty, które spełniają zadane kryteria – nie jest to jednak rozwiązanie doskonałe. Użytkownik może na przykład określić, 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"), określającego charakter poszukiwanych witryn. Prostota tego zagadnienia (i rozwiązania) jest tylko pozorna, ponieważ słowa i całe wyrażenia pojawiają się na różnych stronach i niekoniecznie (lub nie całkowicie) muszą być związane z charakterem konkretnej strony.

Uzasadniona wydaje się więc próba skatalogowania różnych „rodzajów" (ang. 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 [12].

Nie jest możliwe prowadzenie takich operacji w sposób niezautomatyzowany, z uwagi na liczbę stron oraz fakt, że ta liczba stale wzrasta. Konieczne jest opracowanie rozwiązań automatyzujących ten proces i umożliwiających cykliczne jego powtarzanie. Prace prowadzone w tym zakresie obejmowały 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.

Boosting jest 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 – między innymi automatycznej klasyfikacji tekstów [16]. Autorzy zainspirowani tym faktem 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. Należy zwrócić uwagę na fakt, że wiele specyficznych algorytmów stosowanych dla języka angielskiego nie sprawdza się jeśli chodzi o analizę języka polskiego.

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 odmienne 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. Klasy oraz cechy je charakteryzujące (z podziałem na grupy ze względu na ich charakter) zaprezentowano w tabeli 1.

Tabela 1

Charakterystyka i podział stron wg [1]

Grupy cech / Klasy	Zawartość	Forma	Funkcjonalność
Personal Home Page	Informacja o właścicielu strony	Hierarchiczna informacja o powiązanych podtematach	Suwak, adres email, linki do podtematów
FAQ	Pary: pytanie – odpowiedź (Q&A)	Lista	Suwak, opcja wyszukiwania, linki
E-shopping	Lista produktów i usług wraz z opisami	Hierarchia	Suwak, opcja wyszukiwania, adresy email, możliwość zapytań i zamówień online
News	Elementy multimedialne, ankiety, forum, chat	Hierarchia wraz ze znacznikami czasu	Opcja wyszukiwania, adresy email

W artykule [12] zaproponowano podział na 8 gatunków: *link collection*, *help*, *shop*, *portrayal non-private*, *portrayal private*, *article*, *download*, *discussion*. Wykorzystano 1209 stron internetowych, podzielonych na 8 zbiorów (zgodnie z rozpatrywanymi klasami), przy czym liczba stron dla poszczególnych zbiorów była nierówna (wahała się od 123 do 204). Z każdego zbioru losowano po 100 stron i tylko one brały udział w poszczególnych eksperymentach.

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.

Jak widać w tabeli 1, 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. Obecnie trudno o jednoznaczny podział, ze względu na fakt coraz częstszego przenikania się klas, tj. tworzenia witryn, które spełniają wiele funkcji i posiadają cechy pozwalające na przypisanie ich do różnych klas. Nawet wtedy, gdy, jak w przypadku stron typu *FAQ*, wciąż zachowana jest pewnego rodzaju „prostota" takich dokumentów, często są one częścią większych stron (forum, portali itp.). Sytuacja dodatkowo się komplikuje, 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 [4, 12] wykorzystywano znacznie większą liczbę cech, w tym znaki interpunkcyjne, charakterystykę długości strony, różne tagi HTML itp. Prowadzono także eksperymenty na inaczej skonstruowanych podgrupach cech.

Wydaje się, że konieczna jest ekstrakcja dużej liczby, różnych cech. Założenie z góry, że pewne cechy są istotne, a pewne nie, nie wydaje się być prawidłowe. Wątpliwości w tym zakresie rozwiać mogą dopiero przeprowadzone eksperymenty.

W prezentowanym artykule przyjęto podział na 9 klas:

- *Artykuł* – wypowiedź publicystyczna.
- *Blog* – zbiór odrębnych, samodzielnych, uporządkowanych chronologicznie wpisów, których twórcą jest właściciel strony.
- *E-sklep* – sklep internetowy.
- *FAQ* – (ang. Frequently Asked Questions) – zbiory "często zadawanych pytań" i odpowiedzi na nie.
- *Forum* – forma dyskusji, mająca wyodrębnione wątki.
- *Katalog* – moderowany ręcznie zbiór adresów stron internetowych, pogrupowany tematycznie.

- *Portal* – serwis informacyjny dostępny z jednego adresu internetowego, rozbudowany o różnorodne funkcje internetowe.
- *Strona domowa* – prywatna strona internetowa stanowiąca internetową wizytówkę danego użytkownika (właściciela).
- *Strona firmowa* – strona internetowa stanowiąca internetową wizytówkę danej firmy (będącej jej właścicielem).

W eksperymentach wykorzystano (w celach treningowych) 1800 stron, po 200 dla każdej z 9 klas.

3. Cechy opisujące strony internetowe

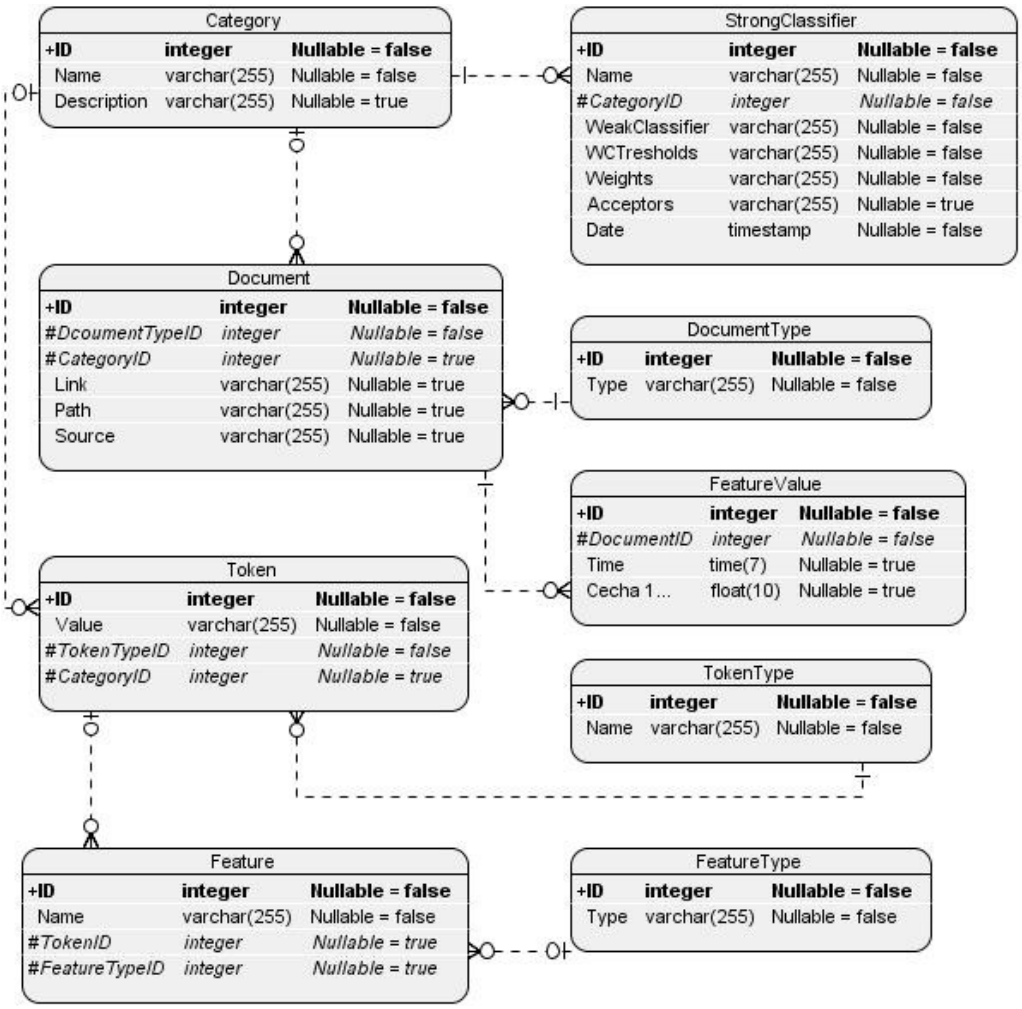
Skuteczna klasyfikacja stron internetowych opiera się na znalezieniu odpowiednich cech, które je charakteryzują. Trudno jest określić z góry, które cechy są na pewno istotne (i okażą się kluczowe w procesie klasyfikacji), a które mają znaczenie marginalne. Wydaje się, że jedyną drogą weryfikacji, które 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), jak i funkcjonalności (skrypty, linki do innych stron) [1].

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

- Cechy tekstowe: statystyki słów kluczowych (zawartych w słownikach zbudowanych dla każdej kategorii), inne statystyki opierające się na słownikach, ogólnych statystykach tekstu, znakach interpunkcyjnych, znakach typograficznych, statystykach 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:
  - stosunek liczby tagów html do ogólnej ilości treści na stronie,

  - stosunek liczby wystąpień sekwencji tagów (tzw. N-gramów) do wszystkich tagów,
  - stosunek ilości kodu skryptowego do pozostałej treści,
  - stosunek ilości kodu skryptowego do ilości 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 ilości 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:
  - związane z formatowaniem – średnie ilości poszczególnych tagów formatujących,
  - związane z obrazami – stosunek ilości 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 ilości plików w różnych formatach do ilości wszystkich plików multimedialnych,
  - związane ze stylem – w tym również występowanie odwołań do zewnętrznych arkuszy CSS.
- Cechy linków do innych stron:
  - 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.

Wszystkie cechy gromadzone są w bazie danych, której struktura zaprezentowana jest na rys. 1. Baza zawiera informacje odnośnie kategorii stron internetowych, cech i ich typów, przykładowe strony (samples), wartości cech dla poszczególnych stron, tokeny kluczowe dla każdej z kategorii oraz ich rodzaje, a także dane potrzebne podczas klasyfikacji. Tabela wartości cech (*FeatureValue*) jest specyficzna – jej kolumny odpowiadają cechom zgromadzonym w tabeli *Feature*. Rozwiązanie to jest efektywne, ponieważ zmniejsza liczbę zapisów do bazy w porównaniu do przypadku, gdy każda wartość cechy miałaby być zapisywana oddzielnie. Na podstawie informacji gromadzonych w prezentowanych tabelach, konstruowane są dane wejściowe dla aplikacji.



Rys. 1. Schemat bazy danych
Fig. 1. Database scheme

4. Aplikacja

Aplikacja służąca do klasyfikacji stron WWW napisana została modułowo. Poszczególne moduły związane są z wydzielonymi etapami procesu uczenia i klasyfikacji witryn. W kolejnych podrozdziałach zostaną one opisane bardziej szczegółowo.

4.1. Przygotowanie słowników kategorii

Na tym etapie analizy stron generowane są (charakterystyczne dla danej klasy decyzyjnej) słowniki zawierające słowa kluczowe. Słowa kluczowe dodawane są 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 sło-

wa, 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 [14].

Moduł nazwany *KewordExtractor* umożliwia użytkownikowi wybór stron poddawanych procesowi wydobywania słów oraz podanie parametrów ekstrakcji. Następnie określa się liczbę i częstotliwości wystąpienia danych słów w dokumentach.

4.2. Ekstrakcja cech

Na podstawie ustalonego zbioru cech strukturalnych, wizualnych, odnośników oraz cech tekstowych poszerzonych o cechy słownikowe (korzystające ze słów kluczowych dla kategorii) moduł *SamplesAnalizer* dokonuje preprocessingu stron, w celu ekstrakcji cech. W wyniku tego z każdą próbką strony WWW zostaje skojarzony wektor cech.

Preprocessing odbywa się kilku etapowo – najpierw analizowany jest kod html dokumentu, czyli cechy związane z występowaniem określonych tagów oraz atrybutów. Na tej postawie generowane są cechy określające zarówno właściwości strukturalne dokumentu (np. liczba wystąpień tagu <table>), cechy wizualne (np. liczba tagów <img>), jak i cechy linków (liczba wystąpień tagów <a> z atrybutem href o takiej samej wartości). W kolejnym etapie odbywa się analiza tekstowa, w której najpierw zajmujemy się analizą znaków, a następnie analizą słów, zgodnie z zasadami omówionymi przy wydobywaniu słów kluczowych dla kategorii.

Końcowym efektem działania modułu SampleAnalyzer jest wektor cech, znormalizowany do przedziału (0, 1).

4.3. Algorytm AdaBoost

Dysponując zbiorem próbek stron WWW oraz ich reprezentacją w postaci wektora cech przystępujemy do budowy klasyfikatorów za pomocą algorytmu AdaBoost. Dla każdej zdefiniowanej kategorii konstruowany jest jeden tzw. silny klasyfikator, będący kombinacją 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 [15]). Autorzy udowodnili, że błąd silnego klasyfikatora końcowego maleje wykładniczo w kierunku zera. Jest to algorytm iteracyjno-uczący, który w kolejnych krokach wybiera najlepsze „słabe" klasyfikatory, opierając się na zbiorze uczącym i dostępnych „słabych" klasyfikatorach. W każdym kolejnym kroku t słabe klasyfikatory h_t 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 α_t określająca jego ważność. Po zakończeniu działania algorytmu (po T krokach) otrzymujemy klasyfikator końcowy H_T, który można obliczyć korzystając ze wzoru:

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

Pseudokod algorytmu

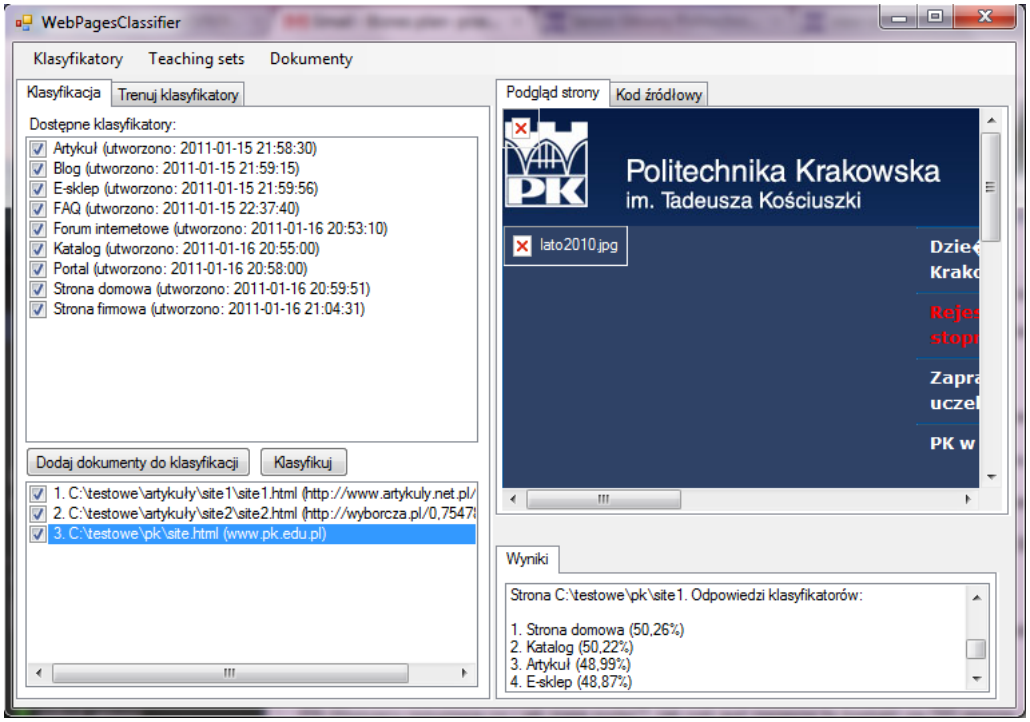
1. Mając zbiór próbek stron (x_i, y_i), …,(x_N,y_N), gdzie y_i= 0,1 odpowiednio dla przykładów negatywnych (strony należące do wszystkich oprócz rozpatrywanej kategorii) i pozytywnych (strony należące do danej kategorii), każdemu elementowi przypisz wagę $d_i^{(1)} = \frac{1}{N}, i = 1,...,N$
2. Dla kroków t=1,…,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 $d_i^{(t+1)} = \frac{d_i^{(t)} \exp\{-\alpha_t y_i h_t(x_i)\}}{Z_t}$, gdzie Z_t jest stałą normalizującą, taką że $\sum_{i=1}^{N} d_i^{(t+1)} = 1$

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

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

4.4. Moduł trenowania klasyfikatorów i klasyfikacji

Moduł nazwany *WebPagesClassifier*, implementujący algorytm *AdaBoost*, jest w stanie wytrenować mocne klasyfikatory dla każdej kategorii, przyjmując jako próbki pozytywne – strony z danej kategorii, a jako negatywne – strony ze wszystkich pozostałych kategorii. Program umożliwia podanie parametrów dla algorytmu *AdaBoost* oraz dokonanie odpowiedniego podziału oznaczonych próbek na zbiór treningowy i testowy. Po wytrenowaniu mocnych klasyfikatorów umożliwia klasyfikację stron ze zbioru testowego i generuje wyniki.



Rys. 2. Aplikacja klasyfikująca
Fig. 2. Classified application

5. Wyniki eksperymentów

W tabeli 2 zamieszczono wyniki eksperymentu polegającego na wytrenowaniu klasyfikatorów z wykorzystaniem 200 stron każdej z 9 klas (1800 stron), a następnie przetestowaniu skuteczności klasyfikacji za pomocą 30 stron dla każdej kategorii (270 stron) niewykorzystywanych w procesie uczenia.

Tabela 2

Tabela krzyżowa skuteczności klasyfikacji stron testowych

Rozpoznanie / Kategorie	Artykuł	Blog	E-sklep	FAQ	Forum internetowe	Katalog	Portal	Strona domowa	Strona firmowa	Niesklasyfikowano
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 internetowe	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%

Z rezultatów eksperymentu zamieszczonych w tabeli 2 wynikają następujące wnioski, dotyczące rozpoznawania poszczególnych kategorii stron:

- Kategoria *Artykuł* jest rozpoznawana z wysoką poprawnością (70%), ze względu jednak na fakt, iż artykuły najczęściej umieszczane są na portalach internetowych, w ramach takich stron rozpoznawane są również elementy *Portali* (13%).
- Kategoria *Blog* jest klasyfikowana z najwyższą w całym zestawieniu skutecznością (76,67%). Wynika to z faktu, że tego typu strony mają charakterystyczną formę i zazwyczaj nie stanowią elementów innych stron. Stąd też błędne sklasyfikowanie do innej kategorii nie przekracza 10%.
- Kategoria *E-sklep* jest rozpoznawana prawidłowo w 63,33%, a błędne przyporządkowania wynikają z faktu, że sklepy internetowe mają obecnie różną formę. Nierzadko stanowią element innych witryn (*portali* – 6,67%, *stron firmowych* – 6,67%) lub zawierają w sobie podstrony innego typu (*FAQ* – 6,67%).
- Kategoria *FAQ* – niska skuteczność prawidłowej klasyfikacji spowodowana jest faktem, iż tego typu strona jest niemal zawsze częścią innej, bardzo złożonej witryny (*Strona firmowa* – 20%, *Portal* – 16,67%). Jednak wciąż, mimo zaledwie 30% skuteczności, żadna inna kategoria nie jest wskazywana częściej dla tego typu strony.
- Kategoria *Forum Internetowe* rozpoznawana jest prawidłowo w 46,67%. Błędne rozpoznania jako FAQ jest łatwe do wytłumaczenia – wynika z częstego łączenia tych kategorii przez twórców stron WWW.
- Kategoria *Katalog* rozpoznawana jest prawidłowo w 50% przypadków.

- Kategoria *Portal* rozpoznawana w 43,33% przypadków, bywa najczęściej klasyfikowana błędnie jako *Artykuł* (20%) lub *Strona Firmowa* (również 20%), co wynika z faktu, że *Portale* są najbardziej złożonymi witrynami, zawierającymi w sobie wiele artykułów i odnośniki do różnych innych strony.
- Kategorie *Strona Domowa* oraz *Strona Firmowa* są często mylone ze sobą. Wynikać to może po części z faktu, że strony tych typów mają podobną formę. Strony obu tych kategorii mają również największy odsetek przypadków niesklasyfikowanych.

W tabeli 2 przedstawiono wyniki drugiego eksperymentu. Tabela obrazuje prawdopodobieństwa dobrego sklasyfikowania próbki, jeśli weźmie się pod uwagę jej wystąpienie w pierwszych dwóch lub 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 3

Skuteczność klasyfikacji stron testowych

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

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.

6. Podsumowanie

W artykule zaprezentowano zastosowanie algorytmu *AdaBoost* do klasyfikacji stron internetowych. Wyniki uzyskane w trakcie eksperymentów są w znacznym stopniu determinowane przyjętymi kategoriami witryn oraz ich liczbą. Wybranie klas, najbardziej różniących

się od siebie, daje szansę na najwyższą skuteczność. W praktyce wybiera się jednak taki zestaw klas, który może być najbardziej przydatny potencjalnemu użytkownikowi, co zdecydowanie zwiększa trudność opracowania systemu. Drugim czynnikiem wpływającym na wyniki jest liczba stron, na jakich przeprowadzano proces tworzenia klasyfikatorów oraz proces ich akwizycji. Konieczne jest zebranie możliwie dużej liczby stron, w porównywalnej ilości dla każdej z rozważanych klas (tak więc zwiększenie liczby klas, na jakie klasyfikujemy zwiększa niezbędną liczbę przykładowych stron). Zbiory tych stron powinny być niezależnie zweryfikowane przez kilka osób, aby stanowiły miarodajny zbiór wzorcowy. Strony w poszczególnych klasach powinny być możliwie różnorodne tak, aby dana kategoria (np. *e-Sklep*) nie zawierała w sobie stron, dotyczących tylko jednego zagadnienia (np. jednego typu produktu). Aspekt doboru właściwego zbioru danych treningowych jest szczególnie ważny w kontekście obranej metody konstrukcji klasyfikatorów za pomocą boostingu. Pomimo tych trudności wyniki uzyskane w prezentowanej pracy stanowią zachętę do dalszego rozwoju podejścia, wykorzystującego algorytm *AdaBoost*. Implementacja innych metod klasyfikacji i wykorzystanie ich dla tych samych zbiorów stron pozwoli na porównanie różnych podejść i ich miarodajną ocenę w przyszłości.

BIBLIOGRAFIA

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, s. 202÷206.
3. Czajkowski K.: Reguły decyzyjne i bazy danych w klasyfikacji stron internetowych, Studia Informatica, Gliwice, Vol. 30, No. 2A(83), 2009, s. 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, 2001, s. 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, s. 1092÷1102.

7. Fernandez, V. F., Unanue, R. M., HerranzS. M., Rubio A. C.: Naive Bayes Web Page Classification with HTML Mark-Up Enrichment. International Multi-Conference on Computing in the Global Information Technology, ICCGI '06, 2006.
8. Xue W., Huang W., Lu Y.: Application of SVM in Web Page Categorization, IEEE International Conference on Granular Computing, 2006, s. 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. Rosmarin A.: The Power of Genre. University of Minneapolis Press, Minneapolis 1985.
11. Yates J., Orlikowski W.: Genres of Organizational Communication: A Structurational Approach to Studying Communication and Media. Academy of Management Review, 17(2), 1992, s. 299÷326.
12. 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, s. 256÷269
13. Roussinov D., Crowston K., Nilan M., Kwasnik B., Cai J., Liu X.: Genre based navigation on the web. In Proceedings of the 34th Hawaii International Conference on System Sciences, 2001.
14. Strona projektu morfologik – http://morfologik.blogspot.com/
15. 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, s. 23÷37.
16. Sebastiani F., Sperduti A., Valdambrini N.: An improved boosting algorithm and its application to automated text categorization. Centre National de la Recherche Scientifique, 2000.

Recenzenci: Dr inż. Dariusz R. Augustyn
Dr hab. inż. Andrzej Chydziński, prof. Pol. Śląskiej

Wpłynęło do Redakcji 28 stycznia 2011 r.

Abstract

One of the main problem for people searching the World Wide Web is a need of browsing a large amount of web pages to find interesting content. Searching based on words and whole

phrases offered by search engines is not always satisfactory. Considered words appear on different pages and do not necessarily (or entirely) must be related to the nature of a specific page.

It seems to be helpful to catalog pages of a desired “genres” and assign them to the appropriate category. Pages belonging to the specific class will be characterized by a similar "style" in terms of form or type of content presentation.

The effectiveness of the classification process depends strongly on the selected classes, their numbers and as far as it is possible on independent features which characterize them. The essential characteristics concern both of the parties (the elements visible to the visitor's side), their structure (types and content of the html tags) and functionality (including scripts, links to other pages).

The study considered nine classes: Article, Blog, e-Shop, FAQ, Forum, Catalog, Portal, Portrayal private, Portrayal non-private. Only Polish pages were collected for training purposes (1,800 pages, 200 for each class).

The construction of genre classifier was based on AdaBoost algorithm. For each defined category a so-called "strong classifier" was constructed as a linear combination of "weak classifiers" (mostly individual features).

Results obtained in experiments are largely determined by the type of page categories and their number. Choosing classes, differing significantly one from another, gives a higher chance for correct classification. The aspect of collection of an appropriate set of training data is particularly important in the context of the chosen method of construction of the classifiers by using boosting. The results obtained in this paper are promising for the further development and employment of boosting methods for genres reliable identification.

Adresy

Tomasz GĄCIARZ: Politechnika Krakowska, Wydział Fizyki, Matematyki i Informatyki, Instytut Teleinformatyki, ul. Warszawska 24, 31-155 Kraków, Polska, tga@pk.edu.pl.

Krzysztof CZAJKOWSKI: Politechnika Krakowska, Wydział Fizyki, Matematyki i Informatyki, Instytut Teleinformatyki, ul. Warszawska 24, 31-155 Kraków, Polska, kc@pk.edu.pl.

Maciej NIEBYLSKI: Politechnika Krakowska, Wydział Fizyki, Matematyki i Informatyki, Instytut Teleinformatyki, ul. Warszawska 24, 31-155 Kraków, Polska, maciej.niebylski@gmail.com.

Ryszard SZAWERNOGA: Politechnika Krakowska, Wydział Fizyki, Matematyki i Informatyki, Instytut Teleinformatyki, ul. Warszawska 24, 31-155 Kraków, Polska, rysiek.szawernoga@gmail.com.