

Technická univerzita v Košiciach
Fakulta elektrotechniky a informatiky
Katedra kybernetiky a umelej inteligencie

Dolovanie webových stránok

Vedúci diplomovej práce:
Marián Mach, M.S., PhD.

Diplomant:
Vladimír Štofanič

Konzultant diplomovej práce:
Marián Mach, M.S., PhD.

Košice 2006

Čestné prehlásenie

Prehlasujem, že som diplomovú prácu vypracoval samostatne s využitím uvedenej odbornej literatúry.

V Košiciach dňa 02.05.2006

.....
vlastnoručný podpis

Na tomto mieste bude vložené zadanie diplomovej práce

Pod'akovanie

Za cenné rady, námety a inšpiráciu by som chcel na tomto mieste poďakovať Mariánovi Machovi, M.S., PhD. z Technickej univerzity k Košiciach, Fakulta Elektrotechniky a informatiky, Katedra Kybernetiky a umelej inteligencie. V neposlednom rade sa chcem poďakovať svojej rodine a známym za ich podporu počas štúdia.

Názov práce : Dolovanie webových stránok

Katedra : Katedra kybernetiky a umelej inteligencie, TU FEI Košice

Autor : Vladimír Štofánik

Vedúci DP : Marián Mach, M.S., Ph.D.

Konzultant DP : Marián Mach, M.S., Ph.D.

Dátum : 02.05.2006

Kľúčové slová : Dolovanie webových stránok, bayesovský klasifikátor, klasifikácia, pedspracovanie textu, Google

Anotácia : Diplomová práca rozoberá a snaží sa riešiť problémy s klasifikáciou webových stránok na základe dostupného textového obsahu, s využitím naivného bayesovského klasifikátora. Celá diplomová práca je rozdelená do niekoľkých tematických celkov, ktoré predstavujú rôzne činnosti úspešnej klasifikácie:

- pedspracovanie textových dát
- výber zástupcov tried
- výber príkladov a kontrapríkladov
- rôzne spôsoby učenia a testovania klasifikátora
- vytváranie dotazov do vyhľadávača Google na získavanie trénovacích príkladov

Program je implementovaný ako internetová aplikácia, ktorá využíva k svojej činnosti pripojenie do siete Internet.

Thesis title : Web pages mining

Department : Department of Cybernetics and Artificial Intelligence, TU
FEI Košice

Author : Vladimír Štofánik

Supervisor : Marian Mach, M.S., Ph.D.

Tutor : Marian Mach, M.S., Ph.D.

Date : 02.05.2006

Keywords : Web mining, naive bayes classifier, classification, spider,
google

Annotation : Graduation these describe and resolve problems with
classification web pages on the basis of available text
content, with using the naive bayes classifier. Graduation
these is split into several thematic units, which include
activities of successful classification.

- text preprocessing
- selecting and reinforcing class representatives
- selecting examples and counterexamples
- different methods of learning and testing
- realize interface to request for system Google to obtain
train examples

Program is implemented as internet application, which
require connect to the Internet network.

Obsah

1.	Úvod.....	1
1.1	Predslov	1
1.2	Formulácia úlohy	2
2.	Dolovanie dát.....	2
2.1	Typy dát.....	2
2.2	Úlohy dolovania dát	2
2.3	Metódy dolovania dát.....	4
2.4	Nástroje dolovania dát.....	5
2.5	Naivný Bayesový klasifikátor.....	6
2.6	Dolovanie z webu.....	7
2.7	Dolovanie z obsahu webu.....	9
2.7.1	Štruktúra HTML dokumentov.....	10
3.	Reprezentácia dokumentov	12
3.1	Vektorový model.....	12
3.2	Ohodnocovanie slov	13
4.	Klasifikácia web stránok	13
4.1	Google	16
4.1.1	Web služba SOAP	17
4.2	Testovanie presnosti klasifikácie	18
4.2.1	Vyhodnotenie presnosti klasifikátora	19
5.	Procesy v dolovaní z obsahu webových stránok	20
5.1	Získavanie obsahu z web stránok	22
5.2	Predspracovanie a čistenie dát	23
5.3	Analýza príznakov	24
5.4	Príklady a kontrapríklady	25
6.	Experimenty	26
6.1	Testovanie bez redukcie množiny príznakov	27
6.2	Testovanie na redukovanej množine príznakov.....	31
6.3	Získavanie príkladov	31
7.	Záver	36
8.	Zoznam použitej literatúry	36
9.	Zoznam príloh.....	37
10.	Zoznam obrázkov a tabuliek.....	37

1. Úvod

1.1 Predslov

Dolovanie dát je smer, ktorý sa v poslednom čase stáva jedným z najdôležitejších činností získavania a vyhľadávania relevantných informácií.

Hromadenie veľkého množstva dát evokuje potrebu ich automatického spracovania, čoho logickým dôsledkom je potreba nových technológií na získanie informácií pre podporu rozhodovania.

V súvislosti s relevantnou oblasťou tejto práce sa často používajú výrazy: dolovanie dát, objavovanie znalostí v databáze resp. skratka KDD. Rôzne zdroje vysvetľujú tieto pojmy odlišne. Tu budú uvedené niektoré z nich.

Pomerne zrozumiteľne je dolovanie dát definované v [1] (Simoudis) podľa ktorého:

Dolovanie dát je proces extrahovania platných dopredu neznámych, zrozumiteľných a využiteľných informácií z rozsiahlych databáz a ich využitie pri závažných obchodných rozhodnutiach.

Trochu iný pohľad je v [2] (Mannila) kde autor uvádza že:

Objavovanie znalostí v databáze sa často nazýva aj dolovanie dát a definuje že jeho cieľom je objavovanie užitočných informácií z veľkých súborov dát. Ďalej zdôrazňuje, že KDD je interaktívny a iteratívny proces s niekoľkými krokmi, pričom dolovanie dát je jeden z nich. Pojmy KDD a dolovanie dát sú v tejto publikácii často nevhodne použité tak, že pôsobia ako ekvivalentné.

Podľa [3] (Hedberg) *KDD je skratka odvodená z knowledge discovery and data mining, čo je dosť nevhodné.*

Pojmovo jasná je definícia podľa [4] (Fayyad a kol.), kde sa uvádza, že objavovanie znalostí v databáze je interaktívny a iteratívny proces s niekoľkými krokmi a dolovanie dát je časťou tohto procesu. Samotný proces KDD definujú ako:

Netriviálny proces rozpoznávania platných, nových (neobvyklých), potenciálne užitočných a jednoznačne zrozumiteľných vzorov z dát.

1.2 Formulácia úlohy

Klasifikácia a dolovanie z obsahu webových stránok, z podmnožiny web stránok firiem zaoberajúcich sa podnikateľskými aktivitami v rôznych priemyselných odvetviach. Úlohou diplomovej práce je navrhnúť a otestovať systém na klasifikáciu textov založeného na využití naivného Bayesovho klasifikátora. Systém musí byť schopný získať obsah z dostupných webových stránok, vyberať zástupcov tried, ktoré predstavujú zvolené priemyselné odvetvia. Stav systému predstavujú proces učenia a proces testovania. Systém je potom otestovaný na podmnožine slovenských webových stránok a je vyhodnotená presnosť štatistickými metódami.

2. Dolovanie dát

2.1 Typy dát

Dáta, z ktorých chceme získať nové znalosti môžu mať rozmanité formy. Vo všeobecnosti ide o numerické a symbolické entity. Numerické atribúty predstavujú číselné údaje, číselné vektory, alebo rozmerné polia čísel. Definičný obor týchto reprezentácií predstavuje spojité alebo diskrétne hodnoty.

Symbolické atribúty popisujú kvalitatívne vlastnosti atribútov (hodnoty atribútu *farba* môžu predstavovať frázy *červená, modrá, zelená*. Definičné obory symbolických atribútov môžu vytvárať prirodzené usporiadanie (*základná škola, stredná škola, vysoká škola, ...*), ide o tzv. ordinálne atribúty alebo v opačnom prípade ide o nominálne atribúty.

2.2 Úlohy dolovania dát

Potreba spracovania veľkého objemu informácií umožnila možnosť vzniku novým spôsobom práce a manipulácie nad dátami. Vznikajú techniky inteligentnej a automatickej transformácie dát na použiteľnú informáciu - data mining (Obr. 1). Data mining (ďalej len DM) znamená proces získavania platných, doposiaľ neznámych a potenciálne použiteľných informácií z databáz.

Medzi hlavné úlohy DM patria:

1. Popisné dolovanie v dátach alebo zovšeobecňovanie, čo predstavuje popis skupiny nájdených príbuzných objektov
2. Prediktívne dolovanie v dátach predstavuje kontrolované učenie, výsledkom ktorého je model pre:
 - klasifikáciu – ak je atribút symbolický
 - predikciu – ak je atribút numerický

Tieto modely sa použijú na určenie vlastností nových objektov.

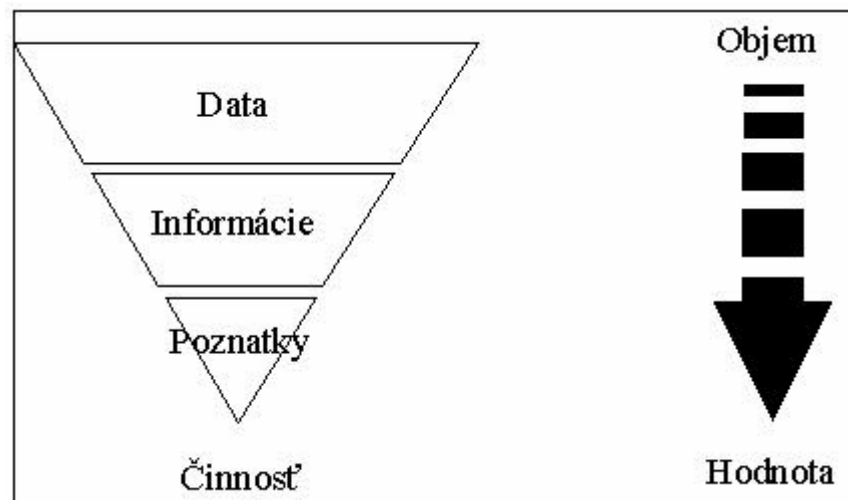
Na DM je obyčajne kladených niekoľko požiadaviek (niektoré môžu byť aj protichodné):

- schopnosť narábať s rôznymi typmi dát (relačné, dočasné, transakčné, priestorové,...), jedna technika nemôže byť z toho dôvodu použiteľná na rôzne typy dát
- efektívnosť a použiteľnosť algoritmu, ktorý by mal byť dostatočne efektívny a mal by byť schopný pracovať na rôznych typoch databáz (algoritmus s exponenciálnou zložitosťou nie je možné použiť)
- použiteľnosť a výrazovosť získaných dát, ktoré musia dostatočne popisovať databázu a musia byť použiteľné pre ďalšiu prácu
- výrazovosť požiadaviek a výsledkov, ktoré predstavujú rôzne znalosti a je potrebné ich aj rôzne interpretovať.
- ochrana súkromia a bezpečnosť dát popisuje, ktoré časti databázy by mali byť pre ktorého používateľa skryté a čo sme schopný danou technikou z databázy získať

Delenie DM možno urobiť podľa rôznych kategórií:

1. typ databázy nad akou pracujem
2. typ znalostí aký sa snažíme získať (asociatívne pravidlá, zhľukovanie, klasifikácia, vývoj,...)
3. akým spôsobom sa snažím získať znalosti (autonómne, dátovo riadené, riadené dotazmi, interaktívne, ...)

Výsledkom DM by mali byť nové poznatky a znalosti, ktoré sme nadobudli zo spracovania. Hierarchicky činnosť DM zobrazuje aj obrázok Obr. 2.1.1.



Obr. 2.1.1: Hierarchia výsledných dát DM

2.3 Metódy dolovania dát

Dáta nad ktorými DM vykonáva činnosti sú rôzneho typu. Nie na všetky typy je možné použiť rovnaké metódy spracovania. Podstatnou úlohou pri DM je potrebné mať jasný cieľ, ktorý chceme pomocou DM dosiahnuť a na ktorého splnenie môžeme použiť viac metód. Z toho vyplýva potreba poznať ich hlavné výhody a mať možnosť porovnať ich výsledky.

Prediktívne modelovanie predstavuje postup hľadania najpravdepodobnejšej hodnoty výstupu na základe známej množiny vstupných hodnôt. Elementárnym príkladom prediktívneho modelovania je napr. hodnotenie rizika úveru v bankovníctve, kedy banka sústreďuje záznamy o všetkých svojich minulých klientoch. Po vytvorení modelu popisujúceho hodnotenie klienta (výstup) na základe informácií o ňom (vstupné dáta), je možné ohodnocovať riziká nových, prichádzajúcich klientov. Používanými technikami pre prediktívne modelovanie sú rôzne typy regresie, neurónové siete a rozhodovacie stromy.

Klasifikácia

Formálna definícia klasifikácie sa vyjadruje pomocou množiny objektov a tried. Nech je daná množina O objektov $o = (o_1, o_2, \dots, o_d)$ kde o_i sú známe hodnoty atribútov A_i , $1 \leq i \leq d$ relevantných vzhľadom na klasifikáciu, a tiež trieda c_j , $c_j \in C = \{c_1, \dots, c_n\}$. Príslušnosť k triede je spravidla vyjadrená hodnotou tzv. cieľového atribútu. Nech D je základný súbor objektov, ktoré je potrebné klasifikovať. Pre každý z objektov v D sú známe hodnoty atribútov A_i , $1 \leq i \leq d$, ale príslušnosť k triede je známa u objektov

z tzv. trénovacej množiny $O \cap D$. Hodnota triedy nie je známa u objektov z $D \setminus O$ (prakticky je známa aj u testovacej množiny aby sme mohli porovnať výsledky testovania).

Klasifikátor je potom funkcia $K, K: D \rightarrow C$.

Regresia

Je štandardná štatistická metóda schopná popisovať stupeň dôležitosti vstupných premenných na výstup. Má prepracovaný odhad chýb modelu a možnosti hľadať závislosti na kombinácii vstupných premenných. Lineárna regresia modeluje cieľový atribút Y (výstup) ako lineárnu funkciu iných, známych atribútov X (vstup), čiže

$$Y = a + b.X$$

pričom regresné koeficienty je a, b je možné určiť napríklad pomocou metódy najmenších štvorcov, ktorá minimalizuje chybu medzi skutočnými dátami a aproximačnou priamkou. Použitie regresie je limitované prácnosťou a časovou náročnosťou vývoja zložitejších modelov.

2.4 Nástroje dolovania dát

Neurónové siete

Predstavujú novú modernú techniku prediktívneho modelovania, s veľkou variabilitou možných modelov a jednoduchosti modifikácie návrhu. Hľadanie parametrov modelu je založené na flexibilnom systéme vnorených funkcií, na druhej strane však nemá zrozumiteľnú interpretáciu.

Rozhodovacie stromy

Patria medzi najpopulárnejšiu formu reprezentácie klasifikátorov. Rozhodovacie stromy predstavuje graf typu strom s vlastnosťami:

- medziľahlý uzol reprezentuje vybraný atribút
- listový uzol reprezentuje triedu
- hrana reprezentuje výsledok testu na atribút, alebo skupinu z nadradeného uzla

Rozhodovací strom sa generuje na základe objektov trénovacej množiny. Nové objekty potom prechádzajú týmto stromom a podľa listového uzla sa zaradzujú do príslušných tried (klasifikácia).

Bayesovská klasifikácia

Bayesove klasifikátory sú štatistické klasifikátory, ktoré predikujú pravdepodobnosti, s ktorými daný príklad patrí do najpravdepodobnejšej triedy. Pritom berú do úvahy podmienené pravdepodobnosti jednotlivých hodnôt atribútov pre triedy do ktorých sa príklad klasifikuje.

Bayesovská teoréma ukazuje ako je možné túto podmienenú pravdepodobnosť určiť:

$$P(H | X) = \frac{P(X | H) \cdot P(H)}{P(X)}$$

Nech X je objekt (povedzme s vlastnosťami červený a guľatý) s neznámym zaradením do triedy C (typ ovocia). H je hypotéza, že X patrí do triedy c_i (jablká). Cieľom klasifikácie Bayesovým klasifikátorom je vlastne určiť podmienenú pravdepodobnosť $P(H|X)$, t.j. v spomenutom prípade pravdepodobnosť toho, že guľatý a červený objekt je jablko. Ide o tzv. aposteriórnu pravdepodobnosť H za podmienky X . Na rozdiel od tejto pravdepodobnosti je $P(H)$ apriórna pravdepodobnosť H . V uvedenom príklade ide o pravdepodobnosť toho, že náš objekt z databázy je jablko, ešte predtým ako zistíme, že je guľatý a červený. Na určenie aposteriórnej pravdepodobnosti $P(H|X)$ je potrebných viac informácií (tzv. background knowledge) než na určenie apriórnej pravdepodobnosti $P(X)$.

Podobne $P(X|H)$ je aposteriórna pravdepodobnosť X za podmienky H , čo predstavuje z príkladu, že ak daný objekt je jablko, nakoľko je pravdepodobné, že je guľaté a červené. $P(X)$ je apriórna pravdepodobnosť toho, že objekt z danej databázy je červený a guľatý.

2.5 Naivný Bayesový klasifikátor

Téma diplomovej práce sa zaoberá využitím bayesovského klasifikátora na klasifikáciu textov extrahovaných z web stránok. Preto je potrebné sa bližšie oboznámiť s touto problematikou a pokúsiť sa čo najlepšie zhodnotiť kvality tohto druhu klasifikácie a jej nasadenia v danom prostredí. Ako už bolo spomenuté bayesovská klasifikácia je vychádza z predpokladu, že efekt, ktorý má hodnota každého atribútu na

danú triedu, nie je ovplyvnený hodnotami ostatných atribútov. Implementácia tohto klasifikátora je použitá aj v programe k tejto práci.

Nech je daná množina objektov O , $o = (o_1, o_2, \dots, o_d)$. Pre každý objekt o sú známe hodnoty atribútov A_i , $1 \leq i \leq d$ relevantných vzhľadom na klasifikáciu, ako aj trieda c_j , $c_j \in C = \{c_1, \dots, c_n\}$. Neznámy príklad $X = (x_1, \dots, x_d)$ bude klasifikovaný do triedy c_i s najväčšou aposteriornou pravdepodobnosťou $P(c_i|X) > P(c_j|X)$ $1 \leq j \leq n$, $i \neq j$.

$$\text{Keďže } P(c_i | X) = \frac{P(X | c_i) \cdot P(c_i)}{P(X)} \text{ a } P(X) \text{ je konštantná pre všetky triedy } c_i,$$

stačí nájsť maximálnu hodnotu výrazu $P(X|c_i) \cdot P(c_i)$ spomedzi všetkých tried c_i . Pravdepodobnosť zaradenia ľubovoľného objektu do triedy c_i je $P(c_i) = N_o^i / N_o$, kde N je počet všetkých príkladov z trénovacej množiny O a N_o^i je počet tých príkladov z O , ktoré patria do triedy c_i .

Ostáva teda určiť pravdepodobnosti $P(X/c_i)$. Tieto pri predpoklade nezávislosti jednotlivých atribútov A_i možno vypočítať nasledovne:

pre kategorické atribúty $P(X | c_i) = \prod_{k=1}^d P(x_k | c_i)$ kde $P(x_k/c_i) = N_o^{i,k} / N_o^i$, pričom N_o^i je počet tých príkladov z O , ktoré patria do triedy c_i a $N_o^{i,k}$ je počet tých z nich, pre ktoré hodnota atribútu $A_k = x_k$.

2.6 Dolovanie z webu

Informácie dostupné na Internete, predstavujú rôzne typy textových a netextových súborov, z ktorých je možné získavať dáta.

Dolovanie z webu je použitie techník dolovania v dátach s cieľom extrahovať a objaviť nové informácie z dokumentov a služieb, ktoré poskytuje sieť Internet.

S rastom počtu webových stránok a služieb priamo klesá úspešnosť hľadania relevantných informácií. Počet dostupných web stránok v súčasnosti predstavuje 11,5 mld., čo je podľa SearchEngineWatch.com¹ len časť celého objemu Internetu. Zvyšok tvorí tzv. Deep Web, ktorý predstavuje:

- obsah databáz dostupných na web stránkach – databázy obsahujú informácie uložené v tabuľkách vytvorených pomocou technológií ako

¹ <http://searchenginewatch.com>

Access, Oracle, SQL Server a DB2, ktoré sú dostupné len cez dotazy (queries). Tento spôsob sa odlišuje od fixných (statických) web stránok, ktorých obsah sa nemení.

- netextové súbory – multimediálne, grafické, software, a dokumenty vo formátoch ako PDF²

Celkový objem Deep Web podľa odhadov presahuje až 500 – násobok objemu statického webu.

Internetové vyhľadávače, ktoré prechádzajú web stránky s cieľom indexovať čo najviac dostupných stránok a dokumentov, sú schopné spracovať len tie ku ktorým smerujú odkazy (hyperlinky³).

Dolovanie z webu používa niekoľko typov dát, s ktorými môžeme pristupovať k objavovaniu znalostí:

- Obsah – spravidla text, v poslednom čase aj grafika
- Štruktúra – organizácia obsahu pomocou HTML⁴ alebo XML⁵ tagov⁶, analýza hyperlinkov
- Používanie – informácie o prístupoch a cestách užívateľov po webe
- Profily návštevníkov – demografické údaje, registračné údaje, nastavenia

Diplomová práca sa snaží využívať práve prvý typ dát získateľných z web stránok. Nakoľko je získaný obsah reprezentovaný ako dokument, v dolovaní sa k tomu aj takto pristupuje a je možné použiť všetky nástroje dolovania v dokumentoch.

Profily návštevníkov a sledovanie ich záujmov sa vo väčšej miere využívajú na komerčných stránkach virtuálnych obchodov, požičovní a cielených reklamných portálov s cieľom poskytnúť návštevníkom dynamicky obsah zodpovedajúci okruhu ich záujmov, pre ktorý daný portál často navštevujú.

² Portable Document Format – od firmy Adobe

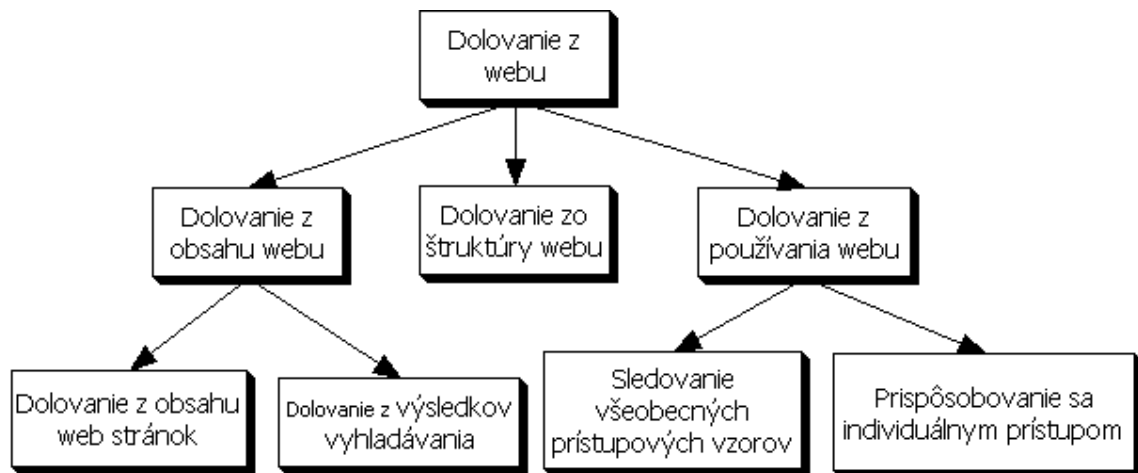
³ odkazy v rámci dokumentu alebo externé odkazy na iné web stránky

⁴ HTML – Hypertext Markup Language, jazyk na tvorbu webových stránok

⁵ XML – Extensible Markup Language, rozšírenie HTML na uchovávanie obsahu a dokumentov

⁶ značka, v rámci HTML alebo XML dokumentu na formátovanie obsahu

Rozdelenie úloh dolovania z webu predstavuje niekoľko samostatných činností v závislosti od potreby typu vstupných dát na ďalšie spracovanie. Obr. 2.4.1 predstavuje v súčasnosti najpoužívanejšie úlohy v dolovaní z webu.



Obr. 2.4.1: Taxonómia úloh dolovania z webu

2.7 Dolovanie z obsahu webu

Dolovanie z obsahu webu predstavuje komplexnú úlohu každého systému, ktorého činnosťou je spracovanie extrahovaných dát z web stránok. Táto úloha predstavuje súbor činností, ktoré je potrebné vykonať pred samotným získaním relevantného obsahu web stránky.

Web stránky obsahujú dáta vo forme textu, obrázkov, audio a video záznamov, a hyperlinkov. Z hľadiska štruktúrovanosti, dolovanie z obsahu môže predstavovať:

- neštruktúrovaný text (čistý text)
- semi – štruktúrovaný text (HTML)
- štruktúrované dáta (XML, DB)

Každá web stránka zobrazovaná v prehliadači obsahuje niekoľko druhov informácií, ktoré spolu tvoria jeden celok. HTML kód ktorý sa prenáša zároveň s relevantným obsahom, formátuje rozmiestnenie a štýly textov, upravuje celkový konečný vzhľad stránky tak ako sa má zobrazit' koncovému používateľovi. Okrem HTML s požiadavkami manipulácie obsahu na strane klienta, a s potlačením nevýhod

tzv. off-line protokolu HTTP⁷ vznikli nové technológie, ktoré boli postupne implementované do internetových prehliadačov a umožnili interaktívne predkladať nové možnosti webových služieb. Patrí tu JavaScript a CSS, ktoré spolu s HTML kódom z pohľadu dolovania dát predstavujú irelevantné informácie, ktoré je potrebné odstrániť. Text určený na ďalšie spracovanie (klasifikácia alebo zhukovanie) musí mať dostatočnú informačnú hodnotu, ktorá zodpovedá významu spracovávaného dokumentu.

2.7.1 Štruktúra HTML dokumentov

Skratka HTML znamená Hypertext Markup Language. Je to jazyk pre tvorbu dokumentov, ktorý definuje vzhľad textu (veľkosť nadpisov, použité písmo, farby, okraje, ...). Jazyk HTML bol špeciálne vyvinutý (a stále sa vyvíja) za účelom publikovania dokumentov na Internete.

Ako už z názvu vyplýva, pôvodne bol jazyk HTML vytvorený pre zobrazovanie textových dokumentov. V relatívne krátkom čase sa jeho pôvodné príkazy boli doplnené o ďalšie, multimediálne prvky (grafika, animácie, hudba). Takto sa jeho hodnota znásobila. HTML jazyk patrí do skupiny značkovacích jazykov (markup languages). Na popis WWW stránok využíva značky (príkazy), ktoré sú od ostatného textu oddelené zátvorkami. Príkazy HTML môžu mať vo všeobecnosti dva tvary:

- párové príkazy sú tvorené otvárajúcim a uzatvárajúcim príkazom, ktoré sa líšia len tým, že uzatvárajúci príkaz obsahuje pred svojim menom znak / (lomítko) `<príkaz> ... </príkaz>`
- nepárové príkazy sú tvorené jedným príkazom `<príkaz>` alebo `<príkaz />`

Zdrojový kód dokumentu HTML je jednoduchý text, ktorý sa dá prehliadať a upravovať v jednoduchom textovom editore. Jazyk HTML je typografický, to znamená, že výsledný dokument iba popisuje, ale jeho interpretácia je ponechaná na cieľový prehliadač (napr. Netscape Navigator, Internet Explorer, Opera, Mozilla, ...).

Hlavným cieľom prehliadačov je zaistiť prístup k zdrojom, ktoré sa môžu nachádzať kdekoľvek na internete. Zdrojom môže byť objekt na internete a môže ním byť HTML dokumenty, obrázky, programy a pod... K jednoznačnej identifikácii týchto

⁷ http – Hypertext Transfer Protocol – protokol na prenos textových dát po sieti

objektov slúži URI (Uniform Resource Identifier) adresa. URL (Uniform Resource Locator) určuje lokáciu, nakoľko sa jeden objekt môže nachádzať na viacerých miestach..

Tieto URL adresy slúžia jednak pri zadávaní adres v prehliadačoch a jednak priamo v HTML dokumentoch.

Časti kompletnej URL adresy:

1. prenosový protokol (napr. http:, ftp:, news:, telnet:)
2. meno serveru, port
3. prístupová cesta
4. meno súboru

Príklad : <http://www.tuke.sk/dokumenty/index.htm>

V tomto prípade sa prehliadač pokúsi otvoriť súbor *index.htm* na serveri www.tuke.sk v adresári *dokumenty*.

Adresa objektu sa nemusí vždy zadávať ako kompletná URL adresa, ale je možné používať i relatívne adresy. Relatívne adresovanie sa obvykle používa v prípadoch, kedy sa odkazuje na zdroje uložené priamo na serveri. Základným adresárom je adresár, v ktorom je uložený aktuálny dokument.

Celý HTML dokument je umiestnený medzi značkami `<HTML>` a `</HTML>`. Špecifikujú, že ide o HTML dokument. Text mimo týchto značiek WWW prehliadač ignoruje.

Dokument pozostáva z hlavičky a tela. Hlavička dokumentu je vymedzená príkazmi `<HEAD>` a `</HEAD>`. V hlavičke sa uvádza názov stránky pomocou značiek `<TITLE>` a `</TITLE>`, ktorý bude zobrazený v záhlaví okna. Telo dokumentu je uzavreté medzi značkami `<BODY>``</BODY>`. Informácie pre vyhľadávače sa umiestňujú do hlavičky dokumentu pomocou nepárovej značky `<META>` s príslušnými atribútmi v tvare:

```
<META name="description" content="stručný popis stránky" />
```

```
<META name="keywords" content="kľúčové slová" />
```

```
<META name="title" content="názov stránok " />
```

Potom štruktúra HTML dokumentu môže vyzeráť aj takto:

```
<HTML>
```

```
  <HEAD>
```

```
<META name="..." content=" ... ">
<TITLE> Názov WWW stránky </TITLE>
</HEAD>
<BODY>
  Obsah stránky
</BODY>
</HTML>
```

Z pohľadu vyhľadávacích robotov je potrebné dodržať tieto základné pravidlá štruktúrovania dokumentov publikovaných na internete z dôvodu lepšej prehľadnosti a dostupnosti obsahu stránok technológiám vyhľadávania a dolovania dát.

3. Reprezentácia dokumentov

Jednotlivé operácie pre spracovanie dokumentov, ku ktorým patrí aj klasifikácia alebo zhlukovanie nad vstupným príznakovým priestorom vyžadujú vhodnú reprezentáciu dokumentov. Každý dokument je tvorený skupinou termov z celkového *univerza* termov. Pod pojmom *term* budeme rozumieť vybranú textovú jednotku.

Z hľadiska štruktúry sa na každý dokument môžeme pozeráť ako na jednoduché textové pole, pri ktorom neberieme do úvahy štruktúru daného dokumentu, alebo ako štruktúrovaný celok tvorený na základe určitých pravidiel dokumentu, podľa ktorých je dokument členený do kapitol, odstavcov, a pod.

3.1 Vektorový model

Vektorový model (Vector Space Model - VSM) je najčastejšou a zároveň najjednoduchšou reprezentáciou textových dokumentov. Príznakový priestor pre túto reprezentáciu je konštruovaný na základe množiny slov, pričom každý príznak korešponduje s textovou jednotkou v dokumente. Za textovú jednotku je považované slovo. Teda dokument sa reprezentuje ako skupina slov bez ohľadu na štruktúru a sémantiku textu.

Daný model vychádza len zo štatistických charakteristík dokumentu, a to najmä z distribúcie pravdepodobnosti jednotlivých slov extrahovaných z dokumentov a dotazov.

Najjednoduchšia reprezentácia dokumentu predstavuje n - rozmerný vektor $d_i = (w_{1i}, w_{2i}, w_{3i}, \dots, w_{ni})$ pričom w_{ij} je funkčná hodnota vyjadrujúca váhu, resp. dôležitosť j - tého slova v i - tom dokumente a n je rozmer indexovej množiny slov, t.j. tých slov, ktoré tvoria príznaky popisu jednotlivých dokumentov.

3.2 Ohodnocovanie slov

Pre váhu w_{ij} slova t_j v dokumente d_i platí: $w_{ij} = F(d_i, t_j)$, kde funkciu F nazývame aj váhová funkcia a jej vyjadrenie nám definuje rôzne spôsoby ohodnocovania slov. Medzi najpoužívanejšie patria:

Binárne – vyjadrenie pomocou booleovskej funkcie: $F: U \times C \rightarrow \{0, 1\}$, kde C je množina dokumentov a U predstavuje univerzum termov. Platí, že $F(d_i, t_j) = 1$ ak sa slovo t_j nachádza aspoň raz v dokumente d_i a 0 ináč. Táto reprezentácia vyjadruje prítomnosť alebo absenciu slova v dokumente. Neurčuje však hľadisko dôležitosti jednotlivých termov.

Frekvencia termov (TF) – berie do úvahy dôležitosť slova vzhľadom na konkrétny dokument, ale nezohľadňuje význam slova v rámci skupiny dokumentov. Je vyjadrená funkciou: $F: U \times C \rightarrow N^+$, kde $F(d_i, t_j) = k$ – počet výskytov termu t_j v dokumente d_i .

TF-IDF koeficient – je jedným z najčastejšie používaných spôsobov ohodnocovania, zohľadňujúcim dôležitosť slova vzhľadom na celú skupinu dokumentov.

$$TF-IDF(i, j) = TF_{ij} \times IDF_j$$

IDF predstavuje inverznú dokumentovú frekvenciu pre ktorú platí výraz:

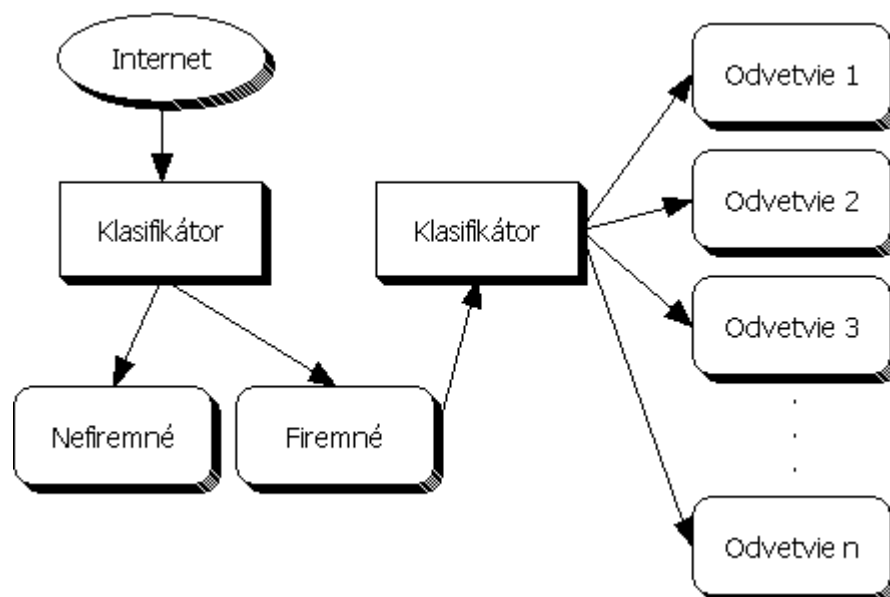
$IDF_j = \log(N / n_j)$, kde N je počet dokumentov a n_j je počet dokumentov obsahujúcich daný term.

4. Klasifikácia web stránok

Klasifikácia web stránok predstavuje samostatnú činnosť systému (klasifikátora) ohodnocovať a zatriedovať nové príklady v podobe získaného obsahu do vopred zvolených tried na základe vopred naučeného popisu množiny tréningových príkladov. Počet tried je vopred pevne definovaný a predstavuje výsledné začleňovanie web stránok na rôznych úrovniach klasifikácie. Diplomová práca rozoberá problematiku

hierarchickej klasifikácie, ktorá predstavuje na 1. úrovni klasifikáciu do dvoch tried (firemných a nefiremných) a následná klasifikácia firemných web stránok do rôznych priemyselných odvetví, Obr. 4.1. Ako klasifikátor bola zvolená metóda Bayesovskej klasifikácie (2.3). Proces klasifikácie pozostáva z dvoch módov:

- proces učenia – klasifikátor získava obsah z dokumentov, ktorý ukladá do slovníka a z nich odhaduje príslušné distribúcie pravdepodobnosti
- proces testovania – klasifikátor na základe naučenej množiny pravdepodobností termov, klasifikuje daný obsah v závislosti od úrovne hierarchie klasifikácie.



Obr. 4.1: Schéma hierarchickej klasifikácie

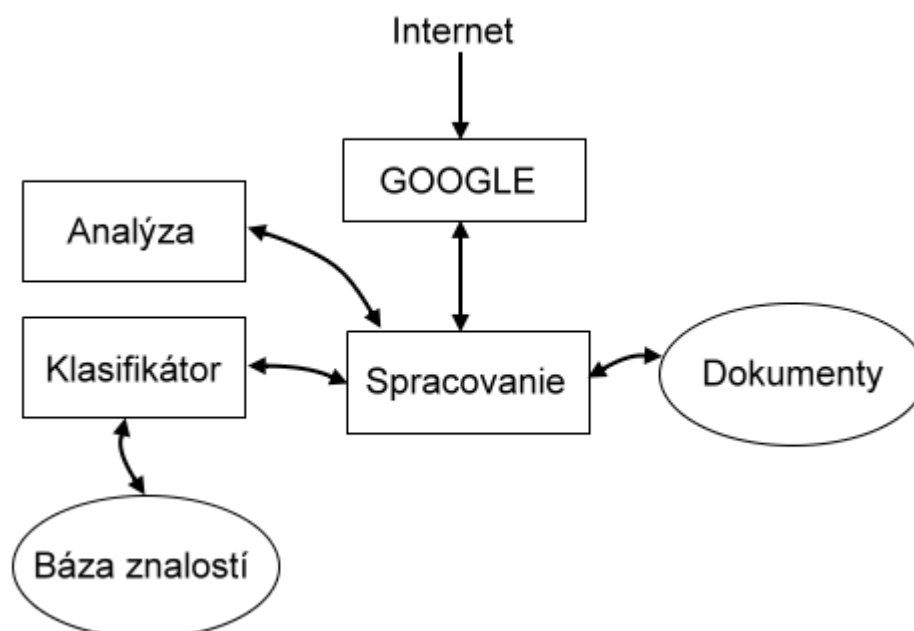
V procese tréningu je hlavnou činnosťou obslužného systému výber zástupcov tried. V tomto smere bola zvolená metóda výberu príkladov a kontra príkladov pomocou dotazovania do systému Google⁸, ktorý na základe vopred formulovanej požiadavky (dotazu) vracia relevantné výsledky, o ktorých predpokladáme, že vrátené dokumenty umožnia prispieť k selekcii zástupcov, klasifikujúcich do správnej triedy. Táto práca chce poukázať niekoľko možností tréningu dokumentov extrahovaných z web stránok, a to v závislosti od vektorovej reprezentácie dokumentu, ktorá v jednom

⁸ <http://www.google.com>

prípade predpokladá výskyt jednotlivých termov v dokumente a v druhom prípade berie do úvahy ich početnosť v rámci dokumentu (ohodnocovanie).

Takto reprezentované dokumenty by mali poskytnúť výsledky vhodné pre ďalšiu diskusiu vhodnosti použitia tohto typu klasifikácie pre konkrétne úlohy. Postupne si rozoberieme všetky mechanizmy v procese dolovania dát z obsahu web stránok. Tieto činnosti ako napr. predspracovanie textu, čistenie dát, operácie nad príznakovým priestorom, majú podľa priebežných výsledkov značný vplyv na výsledky klasifikátora. Preto je dôležité im venovať značnú pozornosť.

Základnú koncepciu navrhnutého systému klasifikácie web stránok a jeho modulov zobrazuje Obr. 4.2:



Obr. 4.2: Základná koncepcia systému klasifikácie

Popis jednotlivých blokov systému klasifikácie:

- **Google** – vracia URL na základe dopytu
- **Analýza** – analyzuje stránku z hľadiska štruktúry, odkazov a doplnkových informácií o stránke
- **Klasifikátor** – klasifikátor, ktorý na základe bázy znalostí (knowledge base) klasifikuje prichádzajúce príklady
- **Spracovanie** – prijíma na vstup stránky zo systému Google, vykoná činnosti spracovania a ukladá ich obsah do vlastnej databázy (Dokumenty) na ďalšie spracovanie

4.1 Google

Na URL adrese <http://www.google.com> sa nachádza jeden z najlepších a najpoužívanejších fulltextových vyhľadávačov súčasnej doby. Vznikol v porovnaní s ostatnými používanými systémami o niečo neskôr, v roku 1997. Vďaka svojmu hodnoteniu web stránok je schopný filtrovať veľké množstvo nerelevantných výsledkov. Má jednoduché a prehľadné rozhranie, ktorého prostredníctvom je denne posielané cez 200 miliónov dotazov. Vo svojich databázach má podľa posledných informácií viac než 5 mld. stránok, 500 mil. obrázkov a 800 mil. diskusných príspevkov. O jeho kvalitách svedčí aj počet užívateľov, ktorých počet prekračuje 70 mil. jedinečných používateľov mesačne a k jeho rozšíriteľnosti prispieva rozhranie v 88 jazykoch. Patrí medzi 10 najpopulárnejších stránok Internetu. Jeho názov bol odvodený zo slova Googol, čo predstavuje matematický termín 1×10^{100} .

Google na ohodnocovanie relevantnosti stránok a usporiadanie výsledkov na dopyt používateľa používa algoritmus PageRank, ktorý využíva štruktúru web liniek pre výpočet hodnotenia kvality. PageRank je v podstate simuláciou náhodného používateľa, ktorý:

- vychádza zo stránky s náhodným URL
- klikne náhodne na niektorú z liniek danej stránky
- po chvíli skočí na stránku s iným náhodným URL

PageRank danej stránky sa zvyšuje tým viac, čím vyšší je PageRank stránok, ktoré sa na danú stránku odkazujú.

$$PR(A) = (1-d) + d * (PR(T_1)/C(T_1) + \dots + PR(T_n)/C(T_n))$$

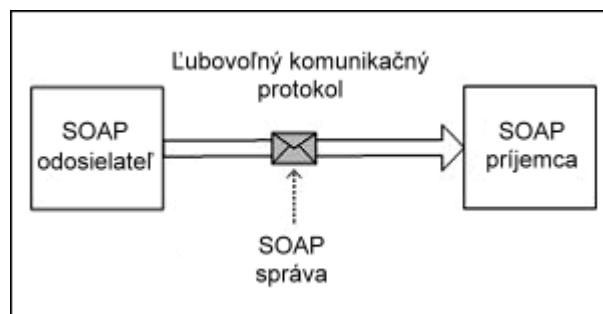
- d – tlmiaci faktor
- T_1, \dots, T_n – stránky odkazujúce na stránku A
- $PR(A)$ – PageRank stránky A
- $PR(T_i)$ – PageRank stránky T_i
- $C(T_i)$ – počet liniek vychádzajúcich zo stránky T_i

Google umožňuje záujemcom o data-mining pomocou rozhrania komunikovať priamo nad jeho databázou s čiastočným obmedzením. Obmedzenie sa týka počtu a dĺžky dotazov. Rozhranie vracia výsledky v podobe URL, zoradené podľa hodnoty PageRanku jednotlivých výsledkov, čo môže pri tréovaní príkladov klasifikátora

značne skomplikovať situáciu. Na popredné pozície sa dostávajú stránky, ktoré majú vhodne štruktúrovaný obsah z hľadiska SEO⁹, a nemusia vždy súvisieť s predstavou používateľa o vrátených výsledkoch. Prístup k databázam vyhľadávača Google je dostupný pomocou web služby SOAP.

4.1.1 Web služba SOAP

SOAP bol pôvodne skratkou pre Simple Object Access Protocol (protokol na jednoduchý prístup k objektom). Predstavoval nástroj na realizáciu DCOM¹⁰ a CORBA (napr. volaní RPC¹¹) cez internet. Pôvodní autori sa zameriavali na „prístup k objektom“, časom sa však od SOAP očakávalo poskytovanie služieb oveľa väčšiemu okruhu záujemcov. Práve preto sa zameranie špecifikácie zmenilo z objektov na všeobecný rámec na komunikáciu XML.



Obr. 4.1.1.1: Odovzdávanie SOAP správ

SOAP je jednoduchý protokol, určený na výmenu štruktúrovaných informácií v decentralizovanom, distribuovanom prostredí. SOAP používa technológie XML na definovanie rozšíriteľného komunikačného rámca poskytujúceho štruktúru správ, ktoré môžu byť vymieňané prostredníctvom množstva základných protokolov. Rámec bol navrhnutý tak, aby bol nezávislý od každého konkrétneho programovacieho modelu a iných špecifických sémantík jednotlivých implementácií. Hlavnou časťou špecifikácie SOAP je komunikačný rámec. Komunikačný rámec SOAP definuje súpravu elementov XML na „zabalenie“ ľubovoľných správ XML na ich výmenu medzi systémami.

⁹ Search Engine Optimization – súbor činností a pravidiel na zlepšenie pozície stránky vo vyhľadávačoch

¹⁰ DCOM, Corba – platformy na budovanie distribuovaných aplikácií

¹¹ RPC – komunikácia vzdialeného volania procedúr

4.2 Testovanie presnosti klasifikácie

Je zrejmé, že na danej trénovacej množine možno v závislosti od použitého algoritmu, resp. v závislosti od nastavenia jeho parametrov, získať mnohé navzájom rôzne klasifikátory. Je preto potrebné vedieť ich navzájom porovnať a zistiť, ktorý z nich je najlepší. Ako najdôležitejšie kritérium porovnania sa používa chyba klasifikácie, t.j. podiel chybné klasifikovaných objektov. Vzniká ale otázka, ktoré dáta sa majú použiť pre odhad chyby klasifikácie. Ak by sa na tento účel použili trénovacie dáta, obvykle by sa získali veľmi nízke hodnoty klasifikačnej chyby, keďže získaný klasifikátor je naučený, a teda vlastne optimalizovaný práve na trénovacie dáta. Tento efekt sa zvykne nazývať preučenie (*overfitting*).

Namiesto toho je možné množinu všetkých známych objektov O rozdeliť na dve podmnožiny.

- trénovaciu množinu, ktorá sa použije v procese budovania klasifikátora (t.j. pre učenie). Často sa zvykne na tento účel vybrať zhruba dve tretiny príkladov z O .
- testovaciu množinu, ktorá sa používa len na odhad chyby klasifikácie pre získaný klasifikátor. Spravidla ide o jednu tretinu príkladov z O .

Táto metóda sa nazýva trénovanie a testovanie. Nemožno ju ale napr. dobre použiť v prípade, keď počet objektov so známou hodnotou triedy je malý a nestačí na naučenie kvalitného klasifikátora, alebo aj vtedy ak nie je celkom jasné delenie na testovaciu a trénovaciu množinu. Vtedy, ale aj vo všeobecnosti pre získanie spoľahlivejšieho odhadu chyby klasifikátora sa zvykne používať tzv. m -násobná krížová validácia (*m-fold cross validation*). Pri m -násobnej krížovej validácii sa množina O rozdelí na m rovnako veľkých podmnožín, z ktorých sa zakaždým použije $m-1$ podmnožín na trénovanie klasifikátora a zvyšná podmnožina potom následne na jeho testovanie. Takto sa získa m rôznych chýb klasifikátora, ktoré sa nakoniec skombinujú pre získanie výsledného odhadu chyby klasifikácie pri použití daného algoritmu a daného nastavenia jeho parametrov. Formálne možno tento postup zapísať nasledovne.

Ak rozdelenie množiny O na podmnožiny nie je náhodné, ale také, aby jednotlivé podmnožiny zachovávali distribúciu jednotlivých tried v každej z podmnožín, ide o tzv. rozvrstvenú násobnú krížovú validáciu (*stratified cross validation*). To znamená, že podiel príkladov z jednotlivých tried je rovnaký tak v

pôvodnej množine O , ako aj v každej z jej podmnožín O_1, \dots, O_m . Vo všeobecnosti sa najčastejšie doporučuje 10-násobná krížová validácia na odhad kvality klasifikátorov.

4.2.1 Vyhodnotenie presnosti klasifikátora

Pri výpočte odhadov kvality môžeme v mnohých prípadoch vychádzať z kontingenčnej tabuľky (Tab. 4.7.1.1), ktorá je generovaná z výsledkov klasifikácie. Každý prvok a_{ij} tabuľky vyjadruje počet príkladov (dokumentov), ktoré boli modelom predikované ako trieda i a v skutočnosti patrili do triedy j . Na diagonále sa teda nachádzajú správne klasifikované príklady.

Kontingenčná tabuľka	Pozitívne príklady	Negatívne príklady
Pozitívne predikované príklady	a (správne pozitívne)	b (nesprávne pozitívne)
Negatívne predikované príklady	c (nesprávne negatívne)	d (správne negatívne)

Tab. 4.2.1.1: Kontingenčná tabuľka pre binárnu klasifikáciu

Na odhad kvality pre problémy binárnej klasifikácie používané v systémoch pre dolovanie dát, v ktorých sa klasifikujú do dvoch tried:

- relevantné používateľovmu dotazu
- irelevantné dotazu používateľa

Koeficient precíznosti ($P = precision$) je definovaný ako pomer relevantných získaných dokumentov ku všetkým získaným dokumentom. Formálne sa to vyjadří ako:

$$P = \frac{a}{a+b}$$

Hodnoty a , b sú získané z kontingenčnej tabuľky výsledkov danej klasifikácie.

Koeficient návratnosti ($R = recall$) je definovaný ako pomer relevantných získaných dokumentov ku všetkým relevantným dokumentom. Analogicky vyjadrenie:

$$R = \frac{a}{a+c}$$

F-metrika je kombináciou predchádzajúcich koeficientov Precision a Recall, kde relatívna dôležitosť každého z nich je vyjadrená hodnotou parametra b :

$$F_b = \frac{(1 + b)^2 \times P \times R}{b^2 \times P + R}$$

kde $b \in (0, \infty)$; pričom ak $b = 0$ potom je metrika F-metrika rovná koeficientu P a ak $b = \infty$ potom je totožná s koeficientom R.

5. Procesy v dolovaní z obsahu webových stránok

V časti (2.6) boli popísané niektoré internetové technológie používané pri formátovaní a práci s obsahom na strane klienta. Tieto časti kódu, ktoré obsahuje každá web stránka je potrebné pred samotným spracovaním odstrániť, aby sa zabránilo nežiaducim termom ovplyvniť výsledky klasifikácie. Na druhej strane, predčasné odstránenie týchto informácií môže viesť k stratám dôležitých údajov a informácií. Navrhnuté riešenie sa v prvom rade sústreďuje na prvotnú analýzu obsahu. Analýza spočíva v niekoľkých činnostiach:

- analýza štruktúry
- analýza odkazov
- analýza doplnkových informácií

Analýza štruktúry mapuje spôsob rozdelenia okna prehliadača. Autori HTML predkladajú dva hlavné spôsoby zobrazenia okna a to: celé okno ako jeden celok (frame) do ktorého sa načítava obsah, alebo okno rozdelené na niekoľko na sebe nezávislých rámcov (frames) (Obr. 4.2.1), z ktorých každý môže zobrazovať iný dokument.



```
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-2">
<meta name="GENERATOR" content="Microsoft FrontPage 3.0">
<title>PVS Computer s.r.o. </title>
</head>
<frameset framespacing="1" border="false" cols="122,*" frameborder="0" rows="*">
  <frame name="left" scrolling="no" noresize target="rtop" src="www/html/left.htm">
  <frame name="rbottom" src="www/html/right.htm" scrolling="auto">
</noframes>
<body bgcolor="#000000">
<p>This page uses frames, but your browser doesn't support them. </p>
</body>
</noframes> </frameset>
</html>
```

Obr. 5.1: Zdrojový kód stránok využívajúcich vnorené okná (frames)

Problém pri druhom spôsobe nastáva ak chceme načítať obsah takéhoto dokumentu ako celku. Väčšina vyhľadávačov sa parsovaniu obsahu stránky zloženej z frames vyhýba z dôvodu náročnosti získať obsah z jednotlivých dokumentov, z ktorých sa daná stránka skladá. Riešenie, ktoré poskytuje aplikácia spočíva v činnosti bloku INDEXER (Obr. 4.2) pri hľadaní všetkých odkazov v rámci stránky a následnom parsovaní obsahu jednotlivých relatívnych URL. Takto je možné extrahovať obsah aj zo stránky s viacerými rozdelenými oknami.

Analýza odkazov umožňuje získať z danej URL nielen obsah úvodnej stránky ale ísť aj do hlbšej úrovne vnorenia. Takto sa do trénovacej množiny dostáva podstatne viac textového obsahu, z ktorého sa dá vybrať väčší počet príznakov na popis daného dokumentu.

Analýza doplnkových informácií sa zaoberá vyhľadávaním doplnkových informácií z hlavičky každej stránky, ktoré približne popisujú obsah web stránky. Nepárové HTML značky <title> a <meta> obsahujú informácie ako názov stránky, krátky popis, a kľúčové slová charakterizujúce zameranie. V skutočnosti veľa autorov stránok na tieto informácie zabúda, alebo vôbec nesúvisia so skutočným obsahom.



Obr. 5.2: Doplnkové informácie

5.1 Získavanie obsahu z web stránok

Proces ukladania obsahu získaného z web stránok predpokladá fyzickú dostupnosť URL adresy z ktorej sa má obsah extrahovať. Servery pri požiadavke na zobrazenie obsahu reaguje správami http protokolu o stave danej URL, ktorá je dostupná ihneď po dotaze (*status code 200*), prípadne je presmerovaná na inú absolútnu, alebo relatívnu adresu (*status code 301*), prípadne je vyvolaná správa chyby (*status code 40x*). Je dôležité sledovať tieto správy http protokolu, aby sme sa vyhli problémom, prípadne načítaniu obsahu z iného zdroja. V praxi existuje mnoho problémov získavania relevantného textového obsahu. Vývojom technológií interaktívnych web portálov sa upustilo od dôležitosti textovej informácie a nastúpili tzv. multimediálne formáty (animované obrázky, FLASH¹²) s primárnym cieľom zaujať a zapôsobiť. Treba si uvedomiť, že stránka zložená z interaktívnych grafických prvkov je možno príjemná na prvý pohľad, ale jej informačná hodnota pre vyhľadávače je prakticky nulová. Vyhľadávače „vidia“ stránky rovnako ako textový prehliadač Lynx¹³, t.j. nevedia parsovať obsah z týchto objektov. Aj keď v súčasnosti už niekoľko robotov prechádzajúce internetom má prostriedky na realizáciu získavania údajov z obrázkov, alebo Flash, je to len vo fázi vývoja (napr. Google dokáže čiastočne Flash prečítať). Asi najväčšou chybou je použitie týchto technológií na navigáciu po stránke. Vyhľadávač v tomto prípade dokáže zaindexovať len prvú stranu. Tzv. Splash page je vstupná stránka, ktorá väčšinou obsahuje len jeden obrázok, alebo Flash animáciu. Po skončení uvítania je návštevník často automaticky presmerovaný. Tieto typy stránok sú v hojnom počte rozšírené po internete a predstavujú zbytočnosť ako pre užívateľov tak aj pre

¹² technológia vyvinutá firmou Macromedia

¹³ <http://lynx.browser.org>

vyhľadávače. Rámce (Obr. 4.2.1) sú dodnes používané ako jeden zo spôsobov tvorby web stránok. Popri tabuľkách a CSS¹⁴ je to základná možnosť ako stránku rozvrhnúť.

Spracovaniu rámcov sa venuje analýza štruktúry (4.2), ktorej výsledkom je získanie všetkých odkazov (hyperlinkov) individuálne z každého definovaného rámca na stránke a ich postupné spracovanie. Alternatívnym textom pri použití Flash animácií a obrázkov na stránkach je očakávaný výstup analýzy doplnkových informácií (4.2), často krát objem týchto informácií je zanedbateľný, a nie je ho možné použiť na relevantné ohodnotenie web stránky.

5.2 Predspracovanie a čistenie dát

Proces čistenia dát predstavuje najdôležitejšiu činnosť v dolovaní dát z webových stránok. Získanie textu z množstva kódu z ktorého sa stránka skladá vyžaduje podrobnú analýzu a činnosti, ktoré musia nasledovať po sebe aby sa zabránilo strate dôležitých dát. Prvotná analýza web stránok (4.2) vyberá z nespracovaného obsahu dôležité dáta, ktoré by sa čistením kódu stratili. Z obsahu sú extrahované odkazy, názov stránky medzi značkami `<TITLE> ... </TITLE>` a doplnkové informácie (Obr. 4.2.2), ktoré by mali predstavovať základnú charakteristiku stránky. V praxi to nie je vždy tak. Tieto dáta sa po procese čistenia pridávajú k extrahovanému obsahu, ktoré potom vstupujú do systému. Je dôležité aby predspracovanie a čistenie dát bolo v procese učenia a testovania rovnaké. Hierarchicky usporiadané činnosti predstavujú základnú koncepciu tohto procesu ktorej výsledkom je pole relevantných termov vstupujúcich do systému:

1. filtrovanie HTML, JavaScript, CSS, a iných zdrojových kódov používaných internetových technológií
2. filtrovanie komentárov
3. filtrovanie HTML entít (` `, `$raquo;`, ...)
4. filtrovanie nežiaducich termov a znakov (*font, color, padding, ...*)
5. filtrovanie na základe dĺžky termov – predpokladáme, že dĺžka relevantného termu je v intervale 3-20 znakov

¹⁴ Cascading Style Sheets – štýly na formátovanie dokumentu

6. filtrovanie termov ktoré sa nachádzajú v tzv. zozname ignorovaných (*ignore list*), a nemajú v procese klasifikácie žiadnu informačnú hodnotu, alebo ich častý výskyt skresľuje výsledky klasifikátora. V slovenskom jazyku to predstavujú zámená (*nami, vami, jemu, tých, tými*), príslovky (*široko, blízko, stále, chutno*), predložky, spojky a častice.
7. filtrovanie termov predstavujúcich len celé slová aj s diakritikou bez špeciálnych znakov

Výsledkom takéhoto procesu sú relevantné príznaky, ktoré svojimi čiastkovými príspevkami zvyšujú pravdepodobnosť zaradenia vstupného príkladu do určenej triedy. Pre proces učenia je potrebný veľký počet relevantných tréningových dokumentov, aby sa zlepšili popisy dokumentov patriacich do jednotlivých tried. Tréningom klasifikátora sa však do bázy znalostí (slovníka) dostáva veľký počet príznakov, ktoré nemusia byť relevantné s obsahom daného dokumentu.

5.3 Analýza príznakov

Analýza príznakov nachádzajúcich sa po tréningu klasifikátora v slovníku má svoj význam pri zlepšovaní celkovej presnosti klasifikácie. Slovník obsahuje veľké množstvo príznakov, ktoré často krát dávajú malé, alebo prakticky žiadne príspevky do tej triedy kam patria. Podľa praktických testov klasifikátora, je presnosť klasifikácie pri takejto množine príznakov podstatne nižšia. Tieto zistenia umožnili vykonanie tejto analýzy nad množinou príznakov. Na druhej strane to má takisto vplyv na rýchlosť a odľahčujú sa týmto systémové nároky klasifikátora. Analýza príznakov rozlišuje dva typy činností:

- redukciu zhodných príznakov, dávajúcich rovnaké príspevky do rôznych tried
- optimalizáciu počtu príznakov v rámci tried

Redukcia zhodných príznakov, prehľadáva slovník klasifikátora, v ktorom sa môže nachádzať niekoľko zhodných príznakov dávajúcich rovnaké príspevky do rôznych tried. Tento stav neutralizuje význam príznaku, ktorý nemá žiadnu informačnú hodnotu. Pri zistení takýchto príznakov sú tieto trvale odstránené zo slovníka. V praxi sa týmto redukuje veľkosť slovníka o 30 - 60%.

Optimalizácia počtu príznakov vyberá z už vopred redukovanej množiny príznakov počet najvýznamnejších v rámci každej triedy. Číslo, ktoré určuje približný počet príznakov pre každú triedu, bolo určené z praktických pokusov pri rôznych mohutnostiach množiny príznakov, tak aby výsledky klasifikátora boli čo najlepšie. V závere sme dospeli k predpokladanej hodnote, predstavujúcu 400 – 500 príznakov pre každú triedu, ktoré môžu byť rovnaké ako príznaky pre iné triedy, líšiace sa počtom výskytu. Nemôže sa stať, aby jeden príznak s rovnakým ohodnotením sa naraz vyskytoval v dvoch alebo viacerých triedach.

5.4 Príklady a kontrapríklady

Kontrapríklady predstavujú dôležitý faktor v procese tréningu klasifikátora, kde plnia úlohu ohraničení výsledného popisu. Vhodným výberom kontrapríkladov sa snažíme zlepšiť aproximáciu krivky rozdeľujúcu pozitívne a negatívne príklady. Myšlienkou výberu kontrapríkladov je analyzovať príznaky a vyberať do tréningovej množiny tie príklady, ktoré obsahujú príznaky triedy T_1 s veľkým príspevkom a príznaky triedy T_2 s malým príspevkom. Takto zaručíme, aby dokumenty v tréningovej množine neboli disjunktné. V procese učenia potom dochádza k zvyšovaniu ohodnotenia silných, alebo vylučovaniu slabých príznakov.

Výberom príkladov získavame nové relevantné dokumenty z obsahu web stránok, ktoré obsahujú vybrané dokumenty. Ide len o proces naplňovania databázy dokumentov, na ktorých potom môžeme robiť podrobnejšie analýzy.

Výber týchto dokumentov je realizovaný zostavením dotazu do systému Google o ktorom predpokladáme, že vráti relevantný výsledok. Dotaz sa zostavuje pomocou vybraných príznakov z analýzy a pomocnými kľúčovými slovami Googla (OR a -) ktorými definujeme, ktoré výrazy má obsahovať hľadaný dokument. Následný výber zo zoznamu vrátených výsledkov pomôže pri naplňovaní a spracovávaní vhodných dokumentov k tréningu klasifikátora.

Dotazy boli zostavované podľa pravidiel a dostupných parametrov na vyhľadávanie, ktorými môžeme ovplyvniť výsledky vyhľadávania. Ide o operátory AND, OR, -. AND operátor zaručí aby sa vo výsledkoch objavili dvojice slov, ktoré spája tento operátor. Napr. dotaz „*slovo1 AND slovo2*“ vracia výsledky s výskytom

obidvoch. Podobne funguje operátor OR ako dotaz „*aspoň jedno zo slov*“. Operátor – zabraňuje aby sa vo výsledku objavil výraz pred ktorým sa tento operátor nachádza.

Keďže rozhranie systému Google kladie obmedzenie na dĺžku dotazu s počtom 10 slov, rozhodli sme sa realizovať výber kontrapríkladov systémom 1:1, a pri výbere príkladov 10: 1, t.j. pri kontrapríkladoch dotaz obsahoval 5 slov, z T_1 a 5 slov z T_2 .

6. Experimenty

Navrhnutý klasifikátor bol otestovaný na množinách dokumentov pri rôznych metódach testovania a rôznych štruktúrach príznakového priestoru. Testovanie klasifikácie dokumentov do 23 tried, pričom pri každom spôsobe je uvedený zoznam parametrov vyhodnotenia kvality testovania (kontingenčná tabuľka, koeficienty precision, recall a F1). V kontingenčných tabuľkách sú pre nedostatok priestoru, triedy označované ako $T_1 - T_{23}$. Zoznam priemyselných odvetví do ktorých sú zatriedované dokumenty predstavuje etalón základných priemyselných odvetví získaný zo slovenských katalógov:

T1 – Automobilový priemysel	T13 – Počítače a internet
T2 – Cestovný ruch a hotelierstvo	T14 - Poľnohospodárstvo
T3 – Chemický priemysel	T15 – Potravinársky priemysel
T4 – Doprava a doručovanie	T16 – Remeselníctvo a výroba
T5 – Drevársky priemysel	T17 – Sklo a keramika
T6 – Elektrotechnický priemysel	T18 – Stavebný priemysel
T7 – Energetický priemysel	T19 – Strojárske priemysel
T8 – Finančné služby	T20 - Suroviny
T9 – Guma a plat	T21 – Textilný priemysel
T10 – Kovový a hutnícky	T22 – Vydavateľstvo a tlač
T11 – Ostatné nezatriedené služby	T23 – Zdravotníctvo
T12 – Celulóza a papier	

6.1 Testovanie bez redukcie množiny príznakov

Prvým typom vyhodnotenia kvality klasifikátora bolo testovanie, na množine 270 dokumentov, ktoré boli náhodne vybrané z katalógov Zoznam.sk¹⁵, Azet.sk¹⁶ a Atlas.sk¹⁷ a naučené na celej množine dokumentov. Podľa teoretických znalostí, by chyba klasifikátora mala byť veľmi nízka. Týmto testovaním chceme poukázať na rozdiel presnosti pri redukovanej (optimalizovanej) a pôvodnej množine príznakov. Testovanie bolo realizované metódami na výskyt príznakov a aj na početnosť. Podmienky za akých bolo testovanie vykonané:

1. učenie a testovanie klasifikátora prebehlo na všetkých dokumentoch
2. slovník obsahoval veľký počet prvkov, z ktorých niektoré svojimi príspevkami prispievali rovnako do rôznych tried
3. porovnáme testovanie metódou výskytu aj početnosti príznakov

Pri metóde na výskyt príznakov je reprezentácia dokumentu vyjadrená ako vektor binárnych hodnôt napr.: $(1, 0, 1, 1, \dots, 0, 0)$, kde 0 na i -tej pozícii ukazuje, že i -té slovo sa v dokumente nevyskytuje, a naopak, 1 na i -tej pozícii vyjadruje prítomnosť i -tého slova v dokumente.

Reprezentácia dokumentu pri metóde početnosti príznakov je vyjadrená ako vektor hodnôt napr.: $(2, 3, 0, \dots, 10, 100)$, ktorého zložky predstavujú počet výskytov i -tého slova v slovníku. Týmto predpokladáme, že ak dokument popisujú slová, ktoré majú vysokú hodnotu početnosti v slovníku, potom patria do množiny zástupcov triedy, do ktorej patria.

Kontingenčné tabuľky pre triedy obsahuje hodnoty z oboch metód testovaní, pričom hodnoty z testovania na početnosť sú označené hviezdičkou (vpravo).

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23
T1	5	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	5	0	0
T2	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0
T3	0	0	3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	6	0	0
T4	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0

¹⁵ <http://www.zoznam.sk/katalog/Priemysel/>

¹⁶ <http://www.azet.sk/katalog/>

¹⁷ http://katalog.atlas.sk/priemysel_a_vyroba/

T5	0	0	0	0	9	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
T6	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	0	1	0	10	0	0
T7	0	0	0	0	1	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T8	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	8	0	0
T9	0	0	0	0	1	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	7	0	0
T10	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0
T11	0	0	0	0	0	0	0	0	0	0	4	1	0	0	0	0	0	0	0	0	5	0	0
T12	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	1	0	0
T13	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0
T14	0	0	0	0	0	0	0	0	0	0	0	0	1	11	0	0	0	0	0	0	0	0	0
T15	1	0	0	0	1	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	9	0	0
T16	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	13	0	0	0	0	0	0	0
T17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0
T18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	1	0	0	0
T19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0
T20	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	7	1	0	0
T21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0
T22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	9	0
T23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	1

Tab. 5.1.1: Kontingenčná tabuľka testu na výskyt

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23
T1	10	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
T2	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
T3	0	0	8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
T4	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
T5	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
T6	0	0	0	0	0	4	0	0	0	1	0	0	0	1	0	0	0	0	1	0	8	0	0
T7	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
T8	0	0	0	0	0	0	0	8	0	0	1	0	0	0	0	0	0	0	0	0	2	0	0
T9	0	0	0	0	1	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	2	0	0
T10	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0
T11	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	2	0	0
T12	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0
T13	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0
T14	0	0	0	0	0	0	0	0	0	1	0	0	0	10	0	0	0	0	0	0	1	0	0
T15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	0	5	0	0
T16	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	13	0	0	0	0	1	0	0
T17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	1	0	0
T18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	1	0	0	0
T19	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0
T20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	2	0	0
T21	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0
T22	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	9	0
T23	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	14	0	1

Tab. 5.1.2: Kontingenčná tabuľka testu na početnosť

	T1	~T1	P = 0.454	R = 0.833	F ₁ = 0.588
T1	5 10*	6 1*	P* = 0.909	R* = 1	F ₁ * = 0.952
~T1	1 0*	258 259*			
	T2	~T2	P = 0.545	R = 1	F ₁ = 0.705
T2	6 10*	5 1*	P* = 0.909	R* = 1	F ₁ * = 0.952
~T2	0 0*	259 259*			
	T3	~T3	P = 0.3	R = 1	F ₁ = 0.462
T3	3 8*	7 2*	P* = 0.8	R* = 1	F ₁ * = 0.889
~T3	0 0*	260 260*			
	T4	~T4	P = 0.545	R = 1	F ₁ = 0.705
T4	6 10*	5 1*	P* = 0.909	R* = 0.909	F ₁ * = 0.909
~T4	0 1*	259 258*			
	T5	~T5	P = 0.9	R = 0.562	F ₁ = 0.692
T5	9 8*	1 2*	P* = 0.8	R* = 0.727	F ₁ * = 0.762
~T5	7 3*	253 257*			
	T6	~T6	P = 0.083	R = 1	F ₁ = 0.153
T6	1 4*	11 11*	P* = 0.267	R* = 1	F ₁ * = 0.421
~T6	0 0*	258 255*			
	T7	~T7	P = 0.917	R = 0.846	F ₁ = 0.88
T7	11 10*	1 2*	P* = 0.833	R* = 1	F ₁ * = 0.909
~T7	2 0*	256 257*			
	T8	~T8	P = 0	R = ∞	F ₁ = ∞
T8	0 8*	11 3*	P* = 0.727	R* = 1	F ₁ * = 0.842
~T8	0 0*	259 259*			
	T9	~T9	P = 0.429	R = 1	F ₁ = 0.6
T9	6 11*	8 3*	P* = 0.786	R* = 1	F ₁ * = 0.88
~T9	0 0*	256 256*			
	T10	~T10	P = 1	R = 0.857	F ₁ = 0.923
T10	12 12*	0 0*	P* = 1	R* = 0.8	F ₁ * = 0.889
~T10	2 3*	256 255*			
	T11	~T11	P = 0.4	R = 0.571	F ₁ = 0.471
T11	4 8*	6 2*	P* = 0.8	R* = 0.727	F ₁ * = 0.761
~T11	3 3*	257 257*			

	T12	~T12	P = 0.909	R = 0.909	F ₁ = 0.909
T12	10 11*	1 0*	P* = 1	R* = 1	F ₁ * = 1
~T12	1 0*	258 259*			
	T13	~T13	P = 1	R = 0.833	F ₁ = 0.909
T13	10 10*	0 0*	P* = 1	R* = 1	F ₁ * = 1
~T13	2 0*	258 260*			
	T14	~T14	P = 0.917	R = 1	F ₁ = 0.957
T14	11 10*	1 2*	P* = 0.833	R* = 0.833	F ₁ * = 0.833
~T14	0 2*	258 256*			
	T15	~T15	P = 0.389	R = 1	F ₁ = 0.56
T15	7 13*	11 5*	P* = 0.722	R* = 1	F ₁ * = 0.839
~T2	0 0*	252 252*			
	T16	~T16	P = 0.867	R = 1	F ₁ = 0.929
T16	13 13*	2 2*	P* = 0.867	R* = 1	F ₁ * = 0.929
~T16	0 0*	255 255*			
	T17	~T17	P = 1	R = 0.909	F ₁ = 0.952
T17	10 9*	0 1*	P* = 0.9	R* = 1	F ₁ * = 0.947
~T17	1 0*	259 260*			
	T18	~T18	P = 0.9	R = 1	F ₁ = 0.947
T18	9 9*	1 1*	P* = 0.9	R* = 1	F ₁ * = 0.947
~T18	0 0*	260 260*			
	T19	~T19	P = 1	R = 0.909	F ₁ = 0.952
T19	10 9*	0 1*	P* = 0.9	R* = 0.9	F ₁ * = 0.9
~T19	1 1*	259 259*			
	T20	~T20	P = 0.7	R = 1	F ₁ = 0.832
T20	7 8*	3 2*	P* = 0.8	R* = 1	F ₁ * = 0.889
~T20	0 0*	260 260*			
	T21	~T21	P = 1	R = 0.111	F ₁ = 0.2
T21	10 9*	0 1*	P* = 0.9	R* = 0.16	F ₁ * = 0.273
~T21	80 47*	180 213*			

	T22	~T22
T22	9 9*	2 2*
~T22	0 0*	259 259*

$$P = 0.818 \quad R = 1 \quad F1 = 0.9$$

$$P^* = 0.818 \quad R^* = 1 \quad F1^* = 0.9$$

	T23	~T23
T23	1 1*	15 15*
~T23	0 0*	254 254*

$$P = 0.063 \quad R = 1 \quad F1 = 0.118$$

$$P^* = 0.063 \quad R^* = 1 \quad F1^* = 0.118$$

Na hlavnej diagonále v kontingenčných tabuľkách (Tab. 5.1.1 a Tab. 5.1.2) sa nachádzajú správne klasifikované príklady.

V teste na výskyt bolo správne klasifikovaných **170** z **270** dokumentov, čo predstavuje presnosť klasifikátora **63.96%**. Pri metóde na výskyt príznakov to predstavovalo **210** z **270** dokumentov s presnosťou **77,78%**.

6.2 Testovanie na redukovanej množine príznakov

Testovanie prebehlo presne podľa postupu ako pri prvom prípade s pôvodnou množinou príznakov. Uvádžam pre prehľadnosť len správne klasifikované príklady, aby sme porovnali výsledky klasifikátora s redukovanou množinou a pôvodnou množinou príznakov.

Po redukcii a optimalizácii veľkosti slovníka predstavoval počet **450** slov popisujúcich každú triedu. Celkovo slovník obsahoval **10 350** fráz. Výsledný počet správne zatriedených dokumentov bol **238**, čo predstavovalo presnosť **88,15%** pri braní do úvahy výskyty príznakov a **221** správne zatriedených dokumentov s presnosťou **81.85%**.

Vysoká presnosť aj vzhľadom na pomerne nízky počet dokumentov pri redukovanej množine príkladov hovorí o význame tejto činnosti, ktorá by mala byť bezprostrednou súčasťou klasifikácie dokumentov na princípe štatistických porovnaní obsahov dokumentov so slovníkmi.

6.3 Získavanie príkladov

Nakoľko testovanie klasifikátora metódou krížovej validácie pri takomto počte dokumentov vykazovalo priemernú presnosť veľmi nízku, približne **23,49%**, bolo potrebné získať nové príklady. Počet dokumentov pri trénovaní klasifikátora krížovou

validáciou v jednom kroku predstavoval len 9 dokumentov. Štatisticky významná hodnota podľa praktických výsledkov sa pohybuje okolo 30. Analýza príznakov vybrala zo 450, priemerne 234 zástupcov, z ktorých sme mohli zostavovať dotaz do vyhľadávača. Náhodným výberom, kombinovaním alebo ručnou manipuláciou sme pomocou tohto rozhrania doplnili do databázy ďalších 226 dokumentov, tak aby počet dokumentov na triedu bol minimálne 20. Tento krok chcel poukázať na zmenu presnosti klasifikácie zvyšujúcim sa počtom tréningových príkladov. Krížová validácia dokáže vyhodnotiť, ktoré príklady zvyšujú presnosť klasifikácie. Príklady, ktoré majú malú informačnú hodnotu je možné z databázy odstrániť a takto experimentovať s dokumentmi.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23
T1	1	0	0	0	0	0	0	0	0	0	0	0	0,1	0,1	0	0,2	0	0	0,1	0	0,1	0	0,1
T2	0	1	0	1	0	0	0	0	0	0,1	0,1	0	0	0	0	0	0	0	0	0,1	0,2	0	0,1
T3	0	0	1	0	0	0	0	0	0	0,6	0	0	0	0	0,1	0	0	0	0	0,2	0	0	0
T4	0	0	0	2	0	0	0	0	0	0,2	0	0	0	0	0	0	0	0	0,1	0	0	0	0
T5	0	0	0	0	1	0	0	0	0	0,7	0	0	0	0	0	0,3	0	0,1	0	0	0,1	0	0
T6	0	0	0	0	0	0	0	0	0	0,3	0	0,1	0	0	0	0,4	0	0	0,2	0	0,3	0	0,1
T7	0	0	0	0	0	0	0	0	0	0,5	0	0,1	0	0	0	0,2	0	0,1	0,1	0,3	0,1	0	0,1
T8	0	0	0	0	0	0	0	1	0	0,1	0	0	0,1	0	0	0	0	0	0,1	0	0,2	0	0,1
T9	0	0	0	0	0	0	0	0	0	0,6	0	0,1	0,2	0	0	0	0	0	0,1	0,2	0	0	0
T10	0	0	0	0	0	0	0	0	0	1,3	0	0	0	0,1	0	0,1	0	0	0,2	0,1	0	0	0
T11	0	0	0	1	0	0	0	0	0	0	0,2	0	0,2	0	0	0,6	0,1	0	0	0	0,1	0	0
T12	0	0	0	0	0	0	0	0	0	0,4	0	1,1	0	0	0	0,4	0,1	0	0	0,1	0,1	0	0
T13	0	0	0	0	0	0	0	0	0	0,2	0,1	0	1,1	0	0,1	0,4	0	0	0	0	0	0	0,1
T14	0	0	0	0	0	0	0	0	0	0,3	0	0,1	0	0,3	0,2	0,4	0	0	0	0,2	0	0,1	0
T15	0	0	0	0	0	0	0	0	0	0,6	0	0	0	0	0,6	0,2	0	0,1	0,1	0	0	0	0
T16	0	0	0	0	0	0	0	0	0	0,6	0	0	0	0	0	0,7	0,1	0,1	0,1	0,2	0,2	0	0
T17	0	0	0	0	0	0	0	0	0	0,4	0	0,1	0	0	0	0,4	1	0	0	0	0,1	0	0
T18	0	0	1	0	0	0	0	0	0	0,7	0	0	0	0	0	0,3	0	0,1	0,1	0	0	0	0
T19	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0,4	0	0	0,2	0	0	0	0
T20	0	0	0	0	0	0	0	0	0	0,7	0	0,1	0	0,1	0	0,3	0,2	0	0,1	0,3	0,1	0	0
T21	0	0	0	0	0	0	0	0	0	0,5	0	0	0	0	0	0,3	0	0	0	0	1,2	0	0
T22	0	0	0	0	0	0	0	0	0	0,1	0	0,4	0,1	0	0	0,2	0	0,1	0	0	0,5	0,4	0
T23	0	0	0	0	0	0	0	0	0	0,3	0	0,1	0,1	0	0,2	0,8	0	0	0	0	0,1	0	0,1

Tab. 6.3.1: KT metódy krížovej validácie na základe početnosti príznakov

	T1	~T1	$P = 0.636$	$R = 0.778$	$F_1 = 0.7$
T1	1,4 0,4	0,8 1,6	$P^* = 0.2$	$R^* = 0.444$	$F_1^* = 0.276$
~T1	0,4 1,5	493,4 493,5			
	T2	~T2	$P = 0.272$	$R = 1$	$F_1 = 0.429$
T2	0,6 0,4	1,6 1,6	$P^* = 0.2$	$R^* = 1$	$F_1^* = 0.333$
~T2	0 0	493,8 494			
	T3	~T3	$P = 0.5$	$R = 0.216$	$F_1 = 0.863$
T3	1,1 0,5	1,1 1,5	$P^* = 0.25$	$R^* = 0.385$	$F_1^* = 0.303$
~T3	4 0,8	489,8 493,2			
	T4	~T4	$P = 1.9$	$R = 0.3$	$F_1 = 0.458$
T4	1,9 0,9	0,3 1,1	$P^* = 0.45$	$R^* = 0.36$	$F_1^* = 0.4$
~T4	4,2 1,6	489,6 492,4			
	T5	~T5	$P = 0.318$	$R = 1$	$F_1 = 0.423$
T5	0,7 0,6	1,5 1,4	$P^* = 0,3$	$R^* = 0,6$	$F_1^* = 0,4$
~T5	0 0,4	493,8 493,6			
	T6	~T6	$P = 0.174$	$R = 0.667$	$F_1 = 0.276$
T6	0,4 0,3	1,9 1,7	$P^* = 0,15$	$R^* = 1$	$F_1^* = 0,261$
~T6	0,2 0	493,5 494			
	T7	~T7	$P = 0.091$	$R = 0.4$	$F_1 = 0.148$
T7	0,2 0,3	2 1,7	$P^* = 0,15$	$R^* = 0,231$	$F_1^* = 0,182$
~T7	0,3 1	493,5 493			
	T8	~T8	$P = 0.546$	$R = 0.705$	$F_1 = 0.615$
T8	1,2 0,6	1 1,4	$P^* = 0,3$	$R^* = 0,6$	$F_1^* = 0,4$
~T8	0,5 0,4	493,3 493,6			
	T9	~T9	$P = 0.136$	$R = 0.5$	$F_1 = 0.215$
T9	0,3 0,1	1,9 1,9	$P^* = 0,05$	$R^* = 0,143$	$F_1^* = 0,074$
~T9	0,3 0,6	493,5 493,4			
	T10	~T10	$P = 0.591$	$R = 0.128$	$F_1 = 0.21$
T10	1,3 0,8	0,9 1,2	$P^* = 0,4$	$R^* = 0,084$	$F_1^* = 0,139$
~T10	8,9 8,7	484,9 485,3			
	T11	~T11	$P = 0.091$	$R = 0.5$	$F_1 = 0.154$
T11	0,2 0,2	2 1,8	$P^* = 0,1$	$R^* = 0,5$	$F_1^* = 0,167$
~T11	0,2 0,2	493,6 493,8			

	T12	~T12	$P = 0.5$	$R = 0.5$	$F_1 = 0.5$
T12	1,1 0,4	1,1 1,6	$P^* = 0,2$	$R^* = 0,308$	$F_1^* = 0,242$
~T12	1,1 0,9	492,7 493,1			
	T13	~T13	$P = 0.5$	$R = 0.579$	$F_1 = 0.537$
T13	1,1 0.6	1,1 1.4	$P^* = 0.3$	$R^* = 0.75$	$F_1^* = 0.429$
~T13	0,8 0.2	493 493.8			
	T14	~T14	$P = 0.136$	$R = 0.5$	$F_1 = 0.214$
T14	0,3 0.4	1,9 1.6	$P^* = 0.2$	$R^* = 0.444$	$F_1^* = 0.276$
~T14	0,3 0.5	493,5 493.5			
	T15	~T15	$P = 0.273$	$R = 0.5$	$F_1 = 0.353$
T15	0,6 0.4	1,6 1.6	$P^* = 0.2$	$R^* = 0.211$	$F_1^* = 0.205$
~T15	0,6 1.5	493,2 492.5			
	T16	~T16	$P = 0.318$	$R = 0.106$	$F_1 = 0.159$
T16	0,7 0.4	1,5 1.6	$P^* = 0.2$	$R^* = 0.065$	$F_1^* = 0.098$
~T16	5,9 5.8	487,9 488.2			
	T17	~T17	$P = 0.455$	$R = 0.667$	$F_1 = 0.54$
T17	1 0.8	1,2 1.2	$P^* = 0.4$	$R^* = 0.235$	$F_1^* = 0.296$
~T17	0,5 2.6	493,3 491.4			
	T18	~T18	$P = 0.045$	$R = 0.167$	$F_1 = 0.071$
T18	0,1 0.2	2,1 1.8	$P^* = 0.1$	$R^* = 0.167$	$F_1^* = 0.125$
~T18	0,5 1	493,3 493			
	T19	~T19	$P = 0.091$	$R = 0.133$	$F_1 = 0.108$
T19	0,2 0.2	2 1.8	$P^* = 0.1$	$R^* = 0.105$	$F_1^* = 0.102$
~T19	1,3 1.7	492,5 492.3			
	T20	~T20	$P = 0.136$	$R = 0.176$	$F_1 = 0.154$
T20	0,3 0,1	1,9 1,9	$P^* = 0.05$	$R^* = 0.067$	$F_1^* = 0.057$
~T20	1,4 1.4	492,4 492.6			
	T21	~T21	$P = 0.546$	$R = 0.352$	$F_1 = 0.429$
T21	1,2 0.7	1 1.3	$P^* = 0.35$	$R^* = 0.146$	$F_1^* = 0.205$
~T21	2,2 4.1	491,6 489.9			

	T22	~T22	P = 0.182	R = 0.8	F ₁ = 0.296
T22	0,4 0.5	1,8 1.5	P* = 0.25	R* = 0.385	F ₁ * = 0.303
~T22	0,1 0.8	493,7 493.2			

	T23	~T23	P = 0.046	R = 0.143	F ₁ = 0.069
T23	0,1 0.1	2,1 1.9	P* = 0.05	R* = 0.067	F ₁ * = 0.057
~T23	0,6 1.4	493,2 492.6			

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23
T1	0	0	0	0	0	0	0	0	0	0,5	0	0	0	0,1	0	0,4	0	0	0,1	0	0	0	0,1
T2	0	0	0	0	0	0	0	0	0	0,2	0	0	0	0,1	0,1	0,2	0,1	0,1	0	0	0,1	0,1	0,1
T3	0	0	1	0	0	0	0	0	0	0,4	0,1	0	0	0,1	0	0,2	0,1	0	0	0	0,4	0	0
T4	0	0	0	1	0	0	0	0	0	0,4	0	0	0	0	0	0,4	0	0	0,1	0	0,2	0	0
T5	0	0	0	0	1	0	0	0	0	0,5	0	0	0	0	0	0,4	0	0	0,1	0,2	0,1	0	0
T6	0	0	0	0	0	0	0	0	0	0,4	0	0	0	0	0	0,2	0,1	0	0,2	0	0,2	0	0,2
T7	0	0	0	0	0	0	0	0	0	0,2	0	0	0	0	0,1	0,3	0	0,2	0	0,1	0,1	0,1	0,2
T8	0	0	0	0	0	0	0	1	0	0,5	0	0	0,1	0,1	0,1	0,1	0	0	0,1	0	0,2	0	0,1
T9	0	0	0	0	0	0	0	0	0	0,4	0	0	0	0	0	0,3	0,4	0,2	0	0,3	0,3	0	0
T10	0	0	0	0	0	0	0	0	0	0,8	0	0	0	0	0	0,3	0,1	0	0,2	0,1	0,2	0	0,1
T11	0	0	0	0	0	0	0	0	0	0,2	0,2	0,1	0,1	0	0,1	0,4	0,2	0	0	0	0,3	0,1	0
T12	0	0	0	0	0	0	0	0	0	0,4	0	0,4	0	0	0	0,1	0,3	0	0,1	0,1	0,3	0,2	0
T13	0	0	0	0	0	0	0	0	0	0,2	0	0,1	0,6	0	0,1	0,3	0,2	0	0	0	0,2	0	0,1
T14	0	0	0	0	0	0	0	0	0	0,3	0	0	0	0,4	0,3	0,2	0	0	0	0	0,2	0	0,1
T15	0	0	0	0	0	0	0	0	0	0,5	0	0	0	0	0,4	0,2	0,1	0,2	0,1	0	0	0,2	0
T16	0	0	0	0	0	0	0	0	0	0,3	0	0	0	0	0,1	0,4	0,3	0,2	0,1	0	0,4	0	0,1
T17	0	0	0	0	0	0	0	0	0	0,5	0	0	0	0	0	0,2	0,8	0	0	0	0,4	0	0
T18	0	0	0	0	0	0	0	0	0	0,6	0	0,1	0	0	0	0,2	0	0,2	0,3	0,1	0	0,1	0,1
T19	0	0	0	0	0	0	0	0	0	0,7	0	0	0	0	0,1	0,1	0	0	0,2	0,2	0,2	0	0,1
T20	0	0	0	0	0	0	0	0	0	0,3	0,1	0,2	0	0	0,1	0,2	0,4	0	0,2	0,1	0	0	0
T21	0	0	0	0	0	0	0	0	0	0,6	0	0,1	0	0	0	0,4	0	0	0	0	0,7	0	0,1
T22	0	0	0	0	0	0	0	0	0	0,3	0	0,3	0	0	0	0,3	0,1	0,1	0,1	0,1	0,2	0,5	0
T23	0	0	0	0	0	0	0	0	0	0,3	0	0	0	0,1	0,4	0,4	0,2	0	0	0,2	0,1	0	0,1

Tab. 6.3.2: KT metódy krížovej validácie na základe výskytu príznakov

Pri metóde na početnosť príznakov bola priemerná úspešnosť klasifikácie **31,95%** pričom najlepšia bola **43,48%**. Výsledky, ako sme predpokladali boli pri druhej metóde asi o polovicu horšie. Testovanie na základe výskytu termov predstavovalo priemernú úspešnosť **21,52%** pri najlepšej **32,6%** úspešnosti.

Výsledky testov potvrdili teóriu lepšej klasifikácie, pokiaľ dokument popisujú nielen jednotlivé príznaky zo slovníka ale aj ich početnosti v dokumentoch. Výsledky takisto ukázali na vhodnosti použitých dokumentov k popisom tried. Ďalším krokom

zlepšovania výsledkov klasifikácie by bolo doplnenie prípadne nahradenie tých dokumentov, ktoré majú malú informačnú hodnotu na popis triedy.

Cieľom experimentov bolo porovnať jednotlivé stavy systému klasifikácie dokumentov a ich chovanie zmenou parametrov u ktorých sme predpokladali ich vplyv na presnosť výsledného umiestňovania do vopred definovaných tried.

Výsledky na generované dotazy do vyhľadávača Google na získavanie príkladov a kontrapríkladov obsahovali vo väčšej miere irelevantné výsledky, a bolo potrebné venovať dostatočnú pozornosť ich skutočnému obsahu. Generované dotazy skladajúce sa z kľúčových slov a parametrov vyhľadávania boli vyhodnocované na strane vyhľadávača, ktorý podľa svojich pravidiel určoval dôležitosť jednotlivých slov v dotaze.. Navyše, vyhľadávač Google relevantnosť stránok, podľa ktorých vracia výsledky, ohodnocuje podľa hodnoty Page Rank každej URL (4.1), čo malo takisto vplyv na výsledky a ako sme aj predpokladali, nie vždy súviseli s požiadavkami. Aplikácia umožňuje tieto irelevantné výsledky blokovať, takže je dostupný väčší priestor na experimentovanie s dotazmi a hľadanie optimálnych kombinácií.

7. Záver

Predpokladané možnosti ďalšieho rozšírenia a vylepšenia ponúka rad nástrojov operácií nad slovníkom, ktorý má podľa experimentov veľký vplyv na presnosť tohto typu klasifikátora. Plnenie slovníka spôsobom zjednocovania termov a porovnávaní voči jednej triede, prípadne kombináciami triedy voči sebe, čiže určitým spôsobom generalizácie. Proces generalizácie ako abstrakcie veľkej množiny dát z databázy relevantných pre danú úlohu predstavuje postup z relatívne nízkej konceptuálnej úrovne na vyššie úrovne.

Samozrejme je potrebné experimentovať aj s inými typmi klasifikátorov a porovnať ich použitie pre rôzne domény, prípadne tieto štatistické metódy na základe čiastkových termov doplniť heuristickými, ktoré budú brať do úvahy aj frázy a ich možné kombinácie .

8. Zoznam použitej literatúry

- [1] Simoudis, E.: Reality Check for Data Mining. IEEE EXPERT, Oct.1996, Vol.11, No.5

- [2] Mannila, H.: Methods and Problems in Data Mining. In the proceedings of International Conference on Database Theory, Afrati, F.- Kolaitis, P., Jan. 1997, Delphi, Springer-Verlag.
- [3] Hedberg, S. R.: Searching for the mother lode: tales of the first data miners. IEEE EXPERT, Oct. 1996, Vol.11, No.5, pp. 4-7.
- [4] FAYYAD, U. M., PIATETSKY - SHAPIRO, G.,SMYTH, P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communication of the ACM, Nov. 1996, Vol. 39, No. 11, pp. 27 – 34.
- [5] Machová, K. : Strojové učenie: Princípy a algoritmy. Elfa, Košice, 2002, ISBN 80-89066-51-8
- [6] McCallum A., Nigam K.: A Comparison of event model for Naive Bayes Text Classification. 1998, [online], Dostupné na internete: <<http://www.cs.cmu.edu/~knigam/papers/multinomial-aaaiws98.pdf>>
- [7] Castegnetto, J., Rawat, H., Schumann, S., Scollo, Ch. Veliath D.:Programujeme PHP profesionálne. Computer Press, Praha, 2001, ISBN 80-7226-310-2
- [8] Musciano, C., Kennedy, B. : HTML a XHTML. Computer Press, Praha, 2000 ISBN 80-7226-407-9
- [9] Paralič, J. : Objavovanie znalostí v databázach. Edícia vedeckých spisov Fakulty elektrotechniky a informatiky, TU Košice, 2003
- [10] Meagher, P. : Implement Bayesian inference using PHP. 2004, [online], Dostupné na internete: <<http://www-128.ibm.com/develeperworks/web/library/wa-bayes1/>>

9. Zoznam príloh

1. CD médium – diplomová práca v elektronickej podobe, prílohy v elektronickej podobe.
2. Používateľská príručka
3. Systémová príručka

10. Zoznam obrázkov a tabuliek

Táto časť obsahuje zoznam všetkých tabuliek a obrázkov v diplomovej práci aj s uvedením čísla strany.

Zoznam obrázkov

Obr. 2.1.1: Hierarchia výsledných dát DM	4
Obr. 2.4.1: Taxonómia úloh dolovania z webu	9
Obr. 4.1: Schéma hierarchickej klasifikácie	14
Obr. 4.2: Základná koncepcia systému klasifikácie.....	15
Obr. 4.1.1.1: Odovzdávanie SOAP správ.....	17
Obr. 5.1: Zdrojový kód stránok využívajúcich vnorené okná (frames)	21
Obr. 5.2: Doplnkové informácie	22

Zoznam tabuliek

Tab. 4.2.1.1: Kontingenčná tabuľka pre binárnu klasifikáciu	19
Tab. 5.1.1: Kontingenčná tabuľka testu na výskyt	28
Tab. 5.1.2: Kontingenčná tabuľka testu na početnosť	28
Tab. 6.3.1: KT metódy krížovej validácie na základe početnosti príznakov	32
Tab. 6.3.2: KT metódy krížovej validácie na základe výskytu príznakov	35