Automatsko označavanje nedostupnih i izmijenjenih sjedišta kataloga www.hr

Krešimir Pripužić, *Marin Vuković*

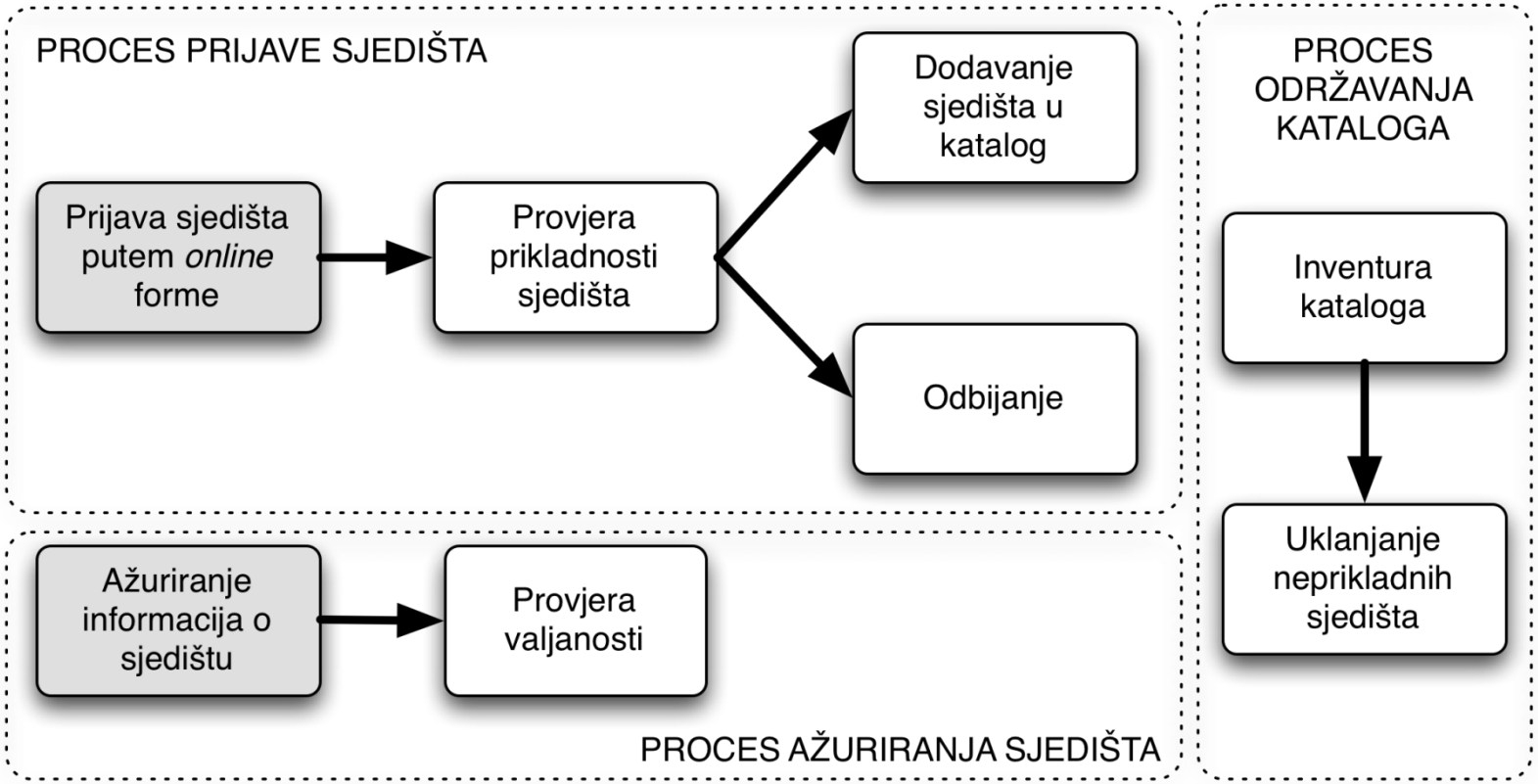
kresimir.pripuzic@fer.hr

Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva

- Projekt je započet 1994. na Zavodu za telekomunikacije, Fakulteta elektrotehnike i računarstva, Sveučilišta u Zagrebu
- Postaje CARNet-ov projekt 1996.
- Trenutno je u katalogizirano preko 25000 sjedišta
- Organizacija kataloga
  - Sjedišta su razvrstana po kategorijama
  - Postoji hijerarhija kategorija
  - Najniže kategorije u hijerarhiji su usko profilirane određenoj djelatnosti, usluzi i slično

Primjer katalogiziranog web sjedišta

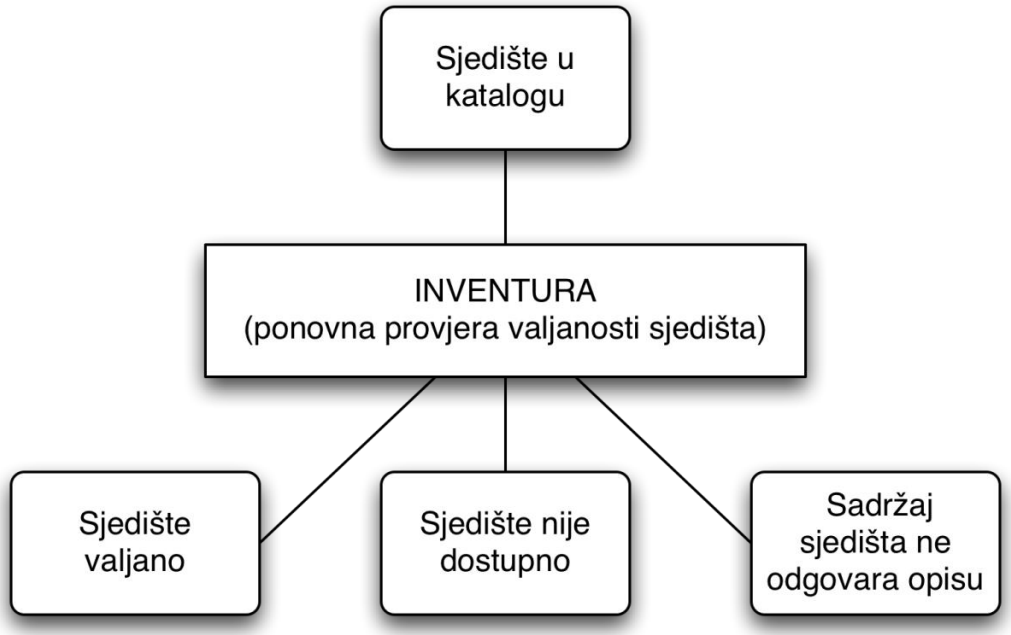
Životni ciklus sjedišta u katalogu



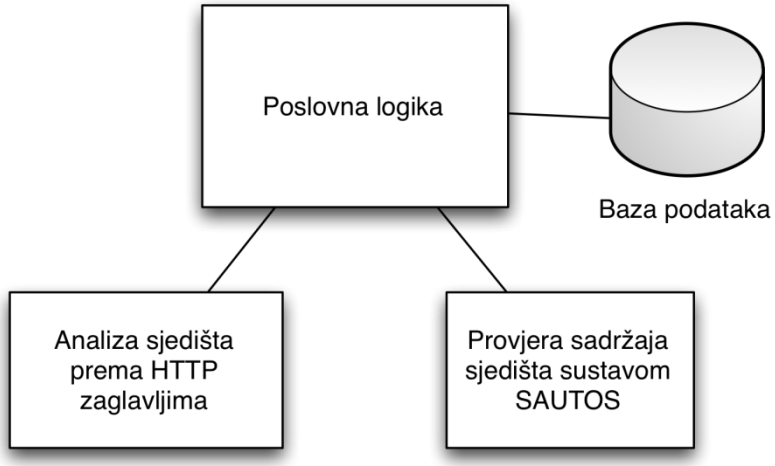
- Prijavu i ažuriranje sjedišta rade sami korisnici
- Provjeru prikladnosti i valjanosti sjedišta te inventuru kataloga radi urednik kataloga

Inventura kataloga

- Provjera valjanosti sjedišta weba u katalogu
- Cilj: čišćenje i ažuriranje kataloga
  - Uklanjanje ukinutih sjedišta weba
  - Uklanjanje neodgovarajućih sjedišta weba
    - ***Štetna sjedišta weba***
    - Sjedišta weba koja više ne odgovaraju svom opisu
    - Neprikladna sjedišta weba
- Ručna inventura kataloga je nemoguća zbog njegove veličine

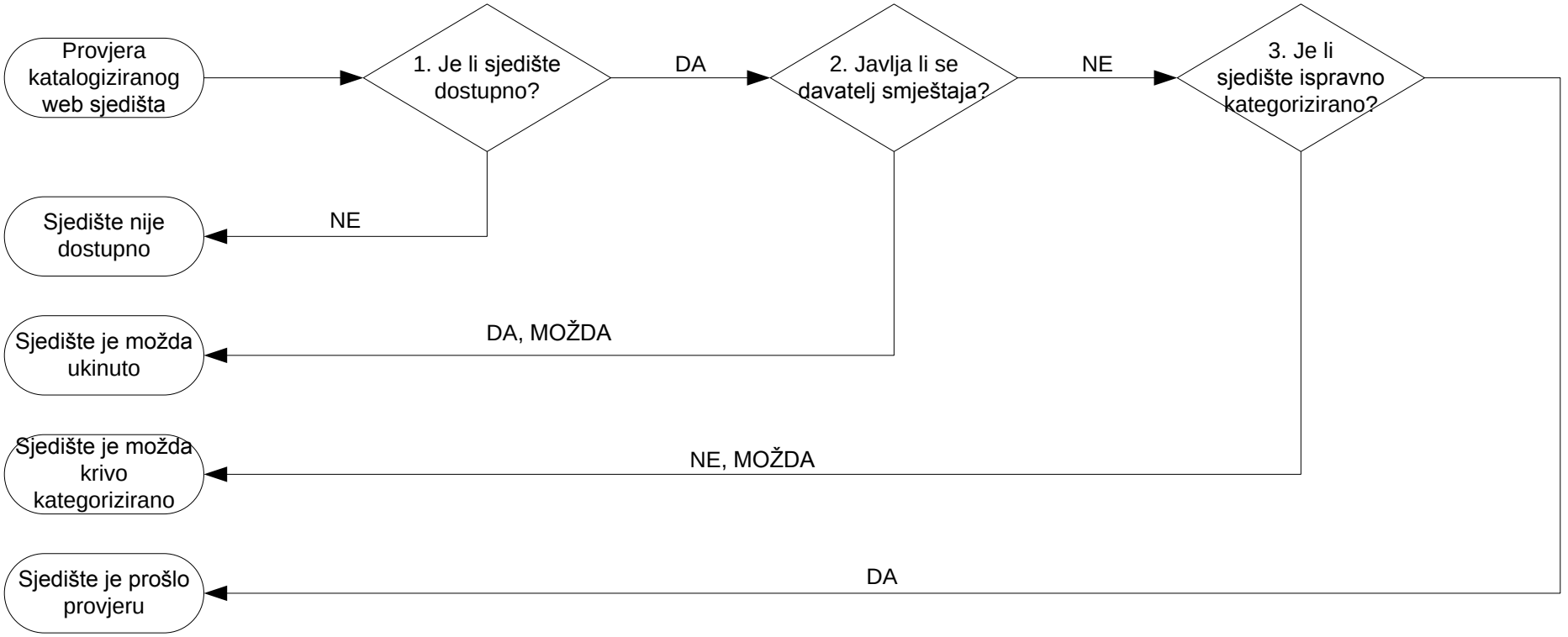


Automatizirano označavanje sjedišta kataloga

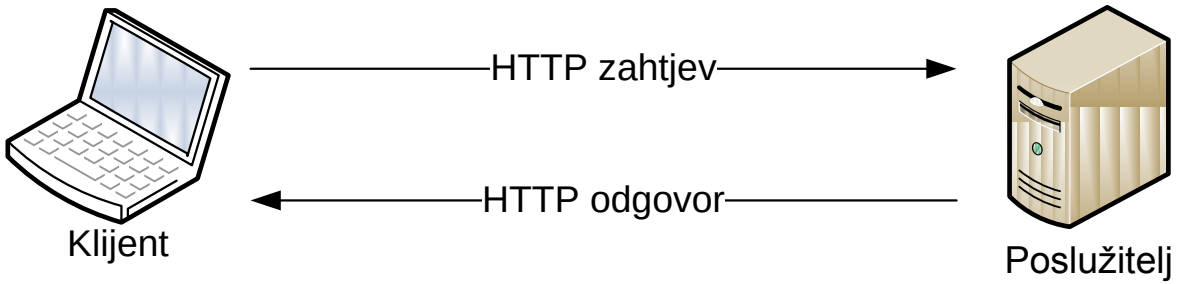


- Proces automatiziranog označavanja se pokreće sa zasebnog računala
  - Provjera sjedišta weba prema statusima u odgovoru poslužitelja sjedišta
  - Provjera sadržaja web sjedišta
- U jednom prolazu se provjerava samo ograničeni skup sjedišta
  - Mogućnost preopterećenja računala s kojeg se vrši provjera

- Sustav za automatizirano utvrđivanje točnosti opisa sjedišta weba u katalogu www.hr



Provjera dostupnosti sjedišta weba

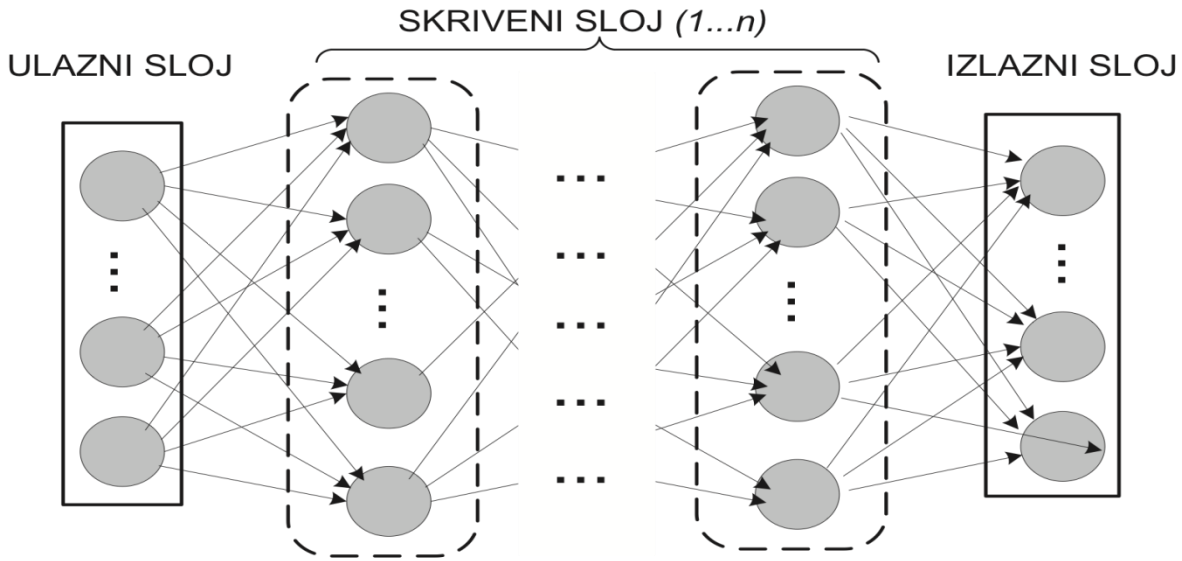


- Za dostupno sjedište poslužitelj odgovara sa statusnim kodom ***200 (OK)***
- Ukoliko postoji problem sa sjedištem poslužitelj odgovara sa statusnim kodovima ***3xx***, ***4xx*** ili ***5xx***
  - ***301 (Moved Permanently)***
  - ***400 (Bad Request)***
  - ***403 (Forbidden)***
  - ***404 (Not Found)***
  - ***503 (Service Unavailable)***
- Lako je automatski identificirati nedostupno sjedište weba
- Obavještava se urednik kataloga

Provjera javljanja davatelja web smještaja

- Umjesto sjedišta javlja se davatelj smještaja s informacijom o nepostojećem sjedištu
- Analiziraju se metapodaci o sjedištu te ključne riječi njegove naslovnice
  - Umjetna neuronska mreža s nadziranim učenjem
- Prepoznaje se javljanje davatelja smještaja
- Označava se sjedište te se obavještava urednik kataloga

Provjera javljanja davatelja web smještaja (2)

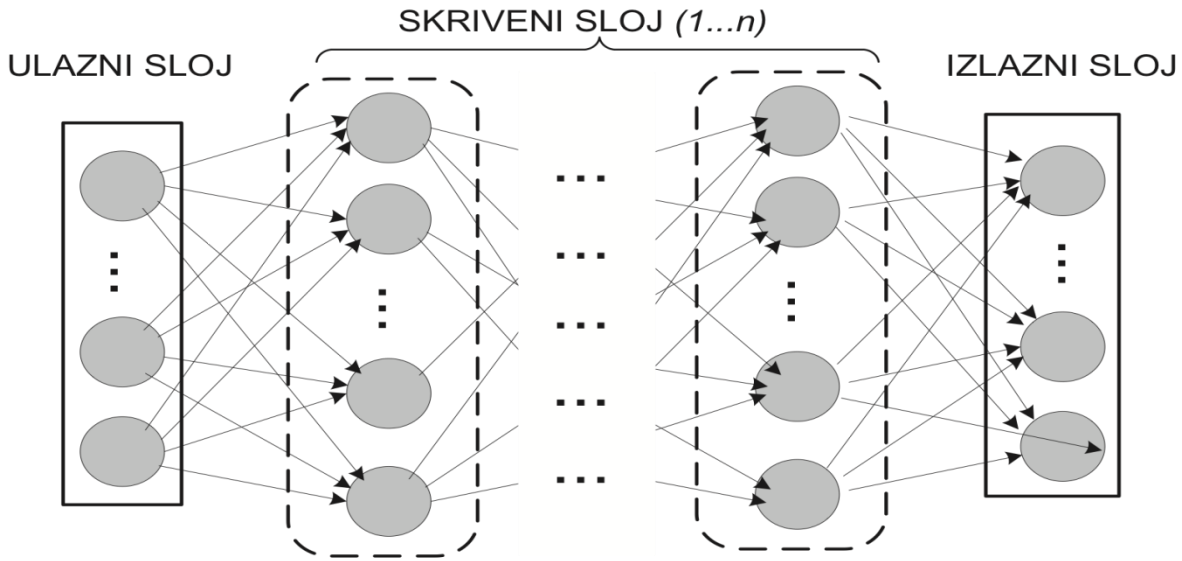


- Ulazni tekst se pretprocesira – vektor ključnih riječi
- Na ulaz umjetne neuronske mreže se dovode vektori
- 2 izlazna neurona
  - mreža procjenjuje vjerojatnost javljanja davatelja web smještaja
- Mreža se uči skupom stranica poznatih davatelja web smještaja i ispravno kategoriziranih sjedišta kataloga

Provjera ispravne kategoriziranosti sjedišta

- Krivo kategorizirano sjedište
  - Promjena djelatnosti tvrtke vlasnika sjedišta weba
- Opis sjedišta u katalogu nije ažuriran
  - Potrebno promijeniti opis sjedišta, njegove ključne riječi i kategoriju
- Metapodaci o sjedištu te ključne riječi njegove naslovnice se uspoređuju s njegovim opisom u katalogu
  - Umjetna neuronska mreža s nadziranim učenjem
- Prepoznaje se krivo kategorizirano sjedište weba
- Označava se sjedište te se obavještava urednik kataloga

Provjera ispravne kategoriziranosti sjedišta (2)



- Ulazni tekst se pretprocesira – vektor ključnih riječi
- Na ulaz umjetne neuronske mreže se dovode vektori
- Broj izlazna neurona odgovara broju krajnjih kategorija u hijerarhiji
  - Mreža procjenjuje kategoriju sjedišta
- Mreža se uči skupom ispravno kategoriziranih sjedišta kataloga

Pretprocesiranje teksta

- Smanjenje dimenzionalnosti ulaznih vektora
- Normalizacija teksta
  - Uklanjanje velikih slova
  - Uklanjanje posebnih znakova
  - Uklanjanje dijakritika
- N-gram tokenizacija teksta
  - N-grami ograničeni na n={3,...8}
  - Kraći n-grami su preopćeniti i imaju prenisku entropiju
  - Dulji n-grami su prespecifični i nisu pogodni zbog generalizacije kod velikog broja različitih sjedišta
- Unaprijed su određeni specifični n-grami na osnovu njihove frekvencije pojavljivanja u kolekciji

Zaključak

- Zbog vjerodostojnosti potrebno je ukloniti nedostupna i krivo kategorizirana web sjedišta iz kataloga
- Drastično se smanjuje obim posla kojeg mora obaviti urednik kataloga
- Poboljšanja sustava
  - Raspodjela procesa provjere na više računala
  - Utvrđivanje optimalnog broja n-grama
  - Nadograditi web sučelje kataloga da bi se olakšalo dojavljivanje krivo kategoriziranih sjedišta weba