Dobývanie znalostí

Vranec Maroš, Lučanský Ján

Zadanie

*Predikcia pozície internetových stránok na kľúčové slovo vo vyhľadávači Google**

* *www.google.cz*

* *site:cz*

Využitie

- Pri **SEO** (Search Engine Optimization)
- Konkrétne pri odhadoch o koľko sa stránka môže posunúť nahor po jej úprave.
  - Z tohto odhadu potom ľahko určiť ako sa zvýši návštevnosť a tým pádom aj predaj daného produktu či služby
  - Čím lepší odhad, tým lepšie (z hľadiska poskytnutia SEO služieb)

Čo ovplyvňuje pozíciu stránky

- **Onpage** faktory
  - Kľúčové slovo v adrese, v title a v Hx elemente stránky
  - Veľkosť stránky
  - Pomer zdrojového kódu ku obsahu
  - Obsah kľúčových slov v obsahu
  - Vek stránky
  - ...
- **Offpage** faktory
  - Počet spätných linkov
  - PageRank, Srank
  - Kvalita spätných linkov
  - ...
- Niektoré atribúty sa dajú zistiť ľahko, niektoré ťažšie a niektoré sa nedajú zistiť vôbec

Kľúčové slová

- Pri získavaní dát sme sa zamerali na kľúčové slová, ktoré súvisia s oborom cestovný ruch
- Vybrali sme **16 jednoslovných** frázi
- Šlo o slová:
  - Dovolená, letenky, exotika, ubytování, cestování, Djerba, Tunisko, Maroko, Řecko, Rhodos, Kréta, Kypr, Mallorca, Bodrum, Egypt, Hurghada,

Zvolené atribúty kľúčového slova

- Počet nalezených odpovedí v Google
  - Číslo
- Konkurenčnosť slova. Informácia z *Google AdWords*
  - Hodnoty 0-5
- Popularita slova
  - Informácia z *Google AdWords*
  - Hodnoty 0-5
- Počet vyhľadávaní slova

  Informácia zo stránky seznam.cz

URL adresy

- Vybrali sme prvých 80 adries, ktoré Google na dané kľúčové slovo zobrazil
- Pre každú z adries sme ako atribút použili:
  - Pozíciu
  - PageRank
  - S-Rank
  - Veľkosť stránky (html dokumentu)
  - Pomer |html dokument| / |obsah|
  - Hustota kľúčového slova
  - Vek domény
  - Prepínače, ktoré indikujú prítomnosť kľúčového slova v názve domény, tagu title a h1

Samotné získanie dát

- Celkovo sme mali 1004 záznamov
- Atribúty kľúčových slov boli získane ručne (len 16 slov)
- URL stránok a atribúty stránok boli získané pomocou voľne prístupných služieb a nami napísaných parserov
- Problémy:
  - Získanie vzorku dát trvalo asi 4 hodiny
  - Parser mal veľa chýb a preto bolo potrebné, ktoré vyšli najavo až po získaní dát

Metóda predikcie

- Použili sme techniku **neurónových sietí**, konkrétne **algoritmus spätného učenia**
- Sieť dostane na vstup atribúty slova a stránky a pozíciu
- Sieť vypočíta výstup a na základe odchýlky od skutočnej hodnoty upraví hodnota synaptických spojov

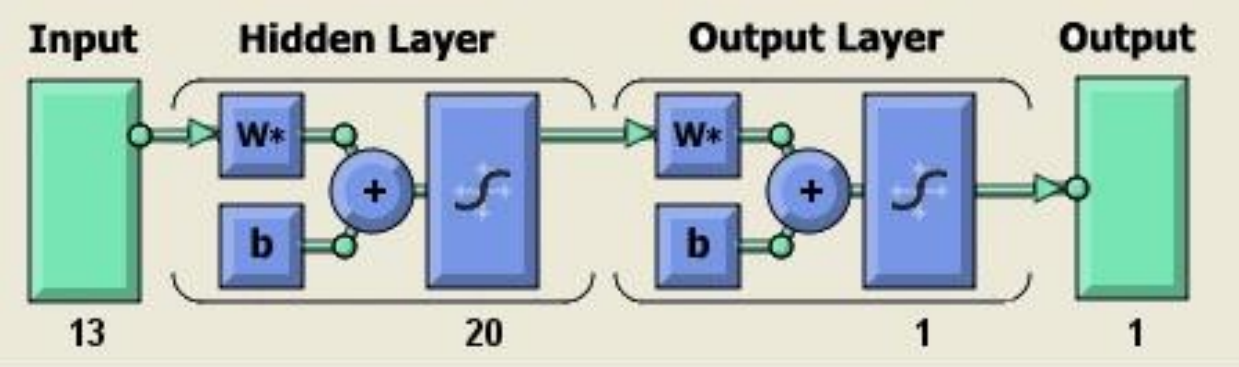
Predpríprava dát

- Normalizácia dát do intervalu [0;1]
- Každý atribút bol vydelený maximálnou hodnotou v jeho stĺpci
- Riadky matice sú náhodne preusporiadané

Rozdelenie dát

- Tréningová skupina – 60% dát
  - Vzory sú predkladané neurónovej sieti počas tréningu a sieť sa im prispôsobuje.
- Validačná skupina – 20% dát
  - Používa sa na kontrolu, či sa ešte sieť zlepšuje. Zabraňuje „preučeniu“.
- Testovacia skupina – 20% dát
  - „Neznáme“ dáta na otestovanie výkonu siete.

Topológia neurónovej sieti

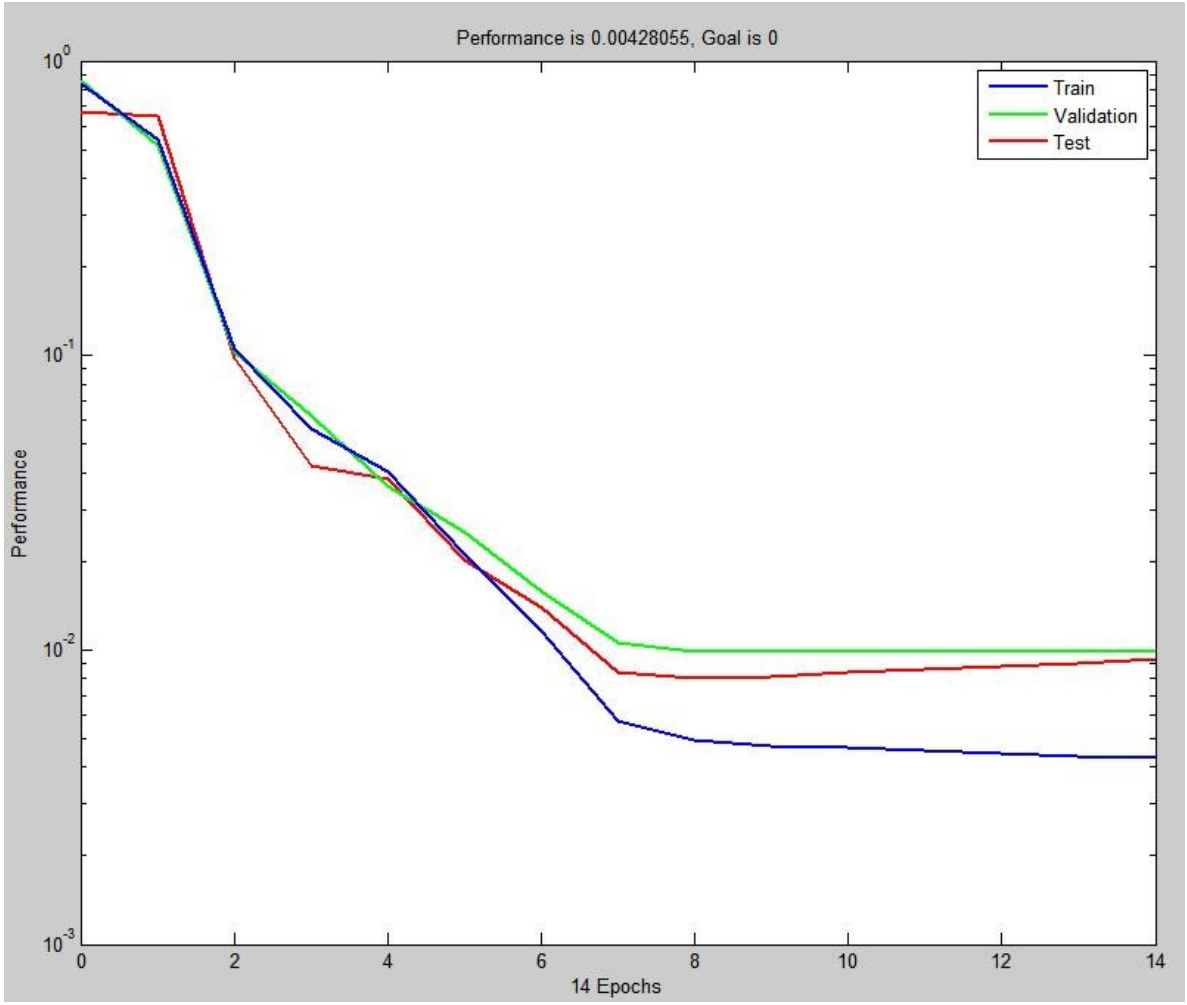


- 13 vstupných parametrov
- 20 skrytých neurónov
- 1 výstupný neurón

Ďalšie parametre siete

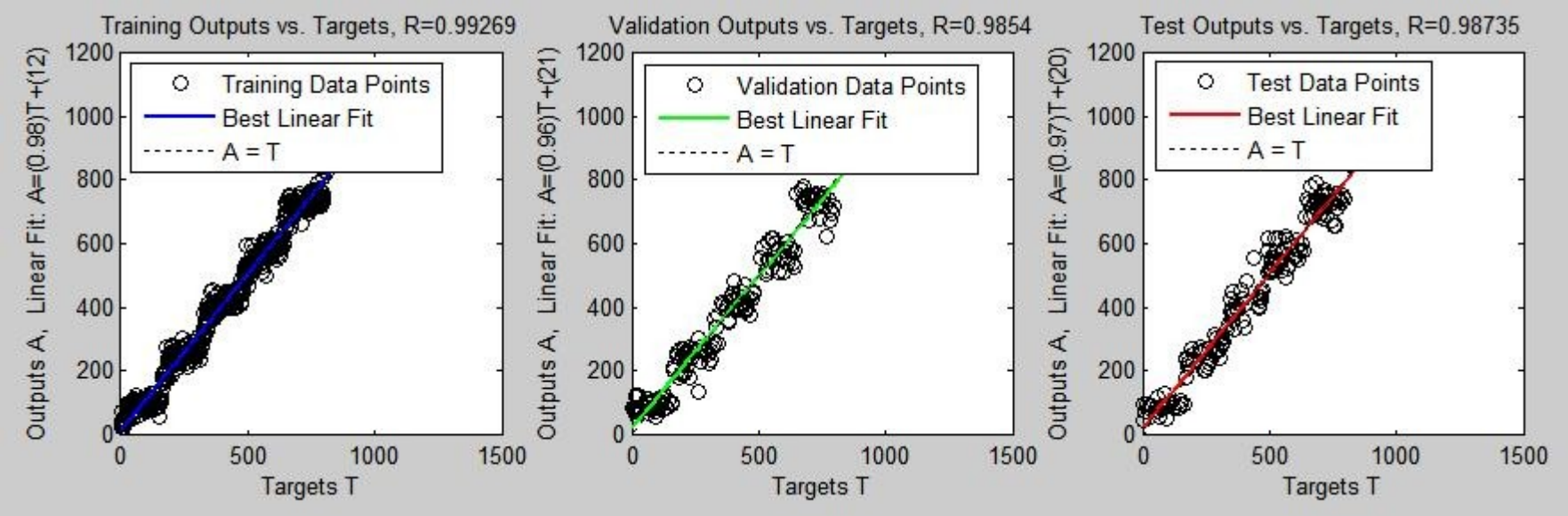
- Prechodové funkcie: logsig
- Trénovacia funkcia: trainlm
- Adaptačná funkcia: trains
- Výkonnostná funkcia: mse

Tréning



- Tréning siete skončil po 14 epochách

Výsledok



- Strednú kvadratickú odchýlku na testovacích dátach sme mali 0,008

Záver

- Vzhľadom k množstve záznamov a sofistikovaným indexovacím nástrojom spoločnosti Google považujeme výsledok za dobrý
- Zaujímavejšie výsledky by prišli po rozšírení množstva atribútov a skúmaných kľúčových slov