Fulltextové vyhľadávanie so štruktúrovaným výsledkom

RNDr. Peter Gurský, PhD. , Patrik Sedlák

Popis práce

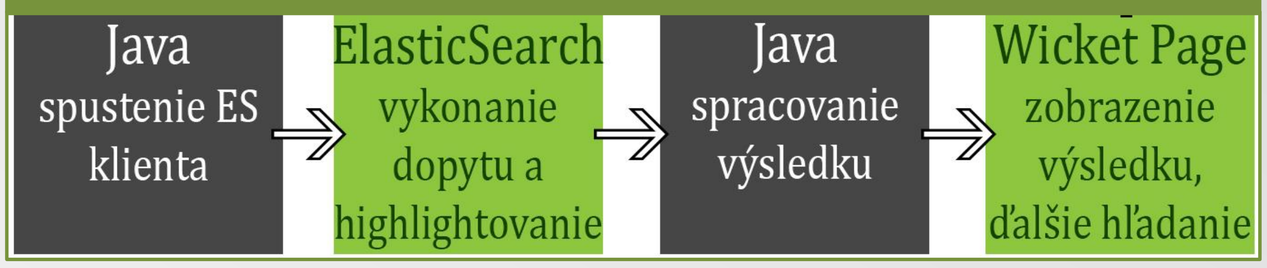
Práca sa zaoberá návrhom a vytvorením metódy fulltextového vyhľadávania nad produktmi internetových obchodov, ktorej výsledkom je okrem zoznamu produktov, ktoré najviac zodpovedajú vstupnému dopytu, aj štruktúrovaná informácia o úspešnosti vyhľadávania podľa domén a atribútov.

Dáta, s akými pracujeme

```
"_source": {
      "id_object": 68660,
      "id_domain": 4,
      "id_source": 17,
      "domain": [
            "Tablety",
            null
      ],
      "attribute_4": "109 eur",
      "attribute_36": "295 g",
      "attribute_37": "3910 mAh",
      "attribute_38": "7 \"",
      "attribute_39": "16 GB",
      "attribute_24": "42318",
      "attribute_40": "1 GB",
      "attribute_41": "áno",
      "attribute_42": "áno",
      "attribute_44": "áno",
      "attribute_32": "189.3 mm",
      "attribute_33": "7 mm",
      "attribute_34": "6 mm",
      "attribute_35": "7950.6 mm3",
      "attribute_26": "Asus MeMO Pad 7 ME176CX-1B046A White",
      "content": "cena 109 eur Hmotnosť Weight 295 g Battery capacity
            Kapacita batérie 3910 mAh Uhlopriečka displeja Display size 7 \"
            Storage Kapacita úložného priestoru 16 GB Local identifier
            Lokálny identifikátor 42318 Memory RAM Pamäť RAM 1 GB GPS áno
            Bluetooth áno WiFi áno Výška 189.3 mm Šírka 7 mm HIbka 6 mm Objem
            7950.6 mm3 Názov Name Asus MeMO Pad 7 ME176CX-1B046A White"
}
```

elasticsearch.

Schéma riešenia



Navrhnuté riešenie

- Pre vyhľadávanie využívame fulltextový engin Elasticsearch, ktorý vie vyhľadáva nad JSON doumentami
- Dáta uložené v Elasticsearch sú štruktúrované, každému produktu zodpovedá jeden JSON dokument
- Ku každému produktu evidujeme pole *“content“*, ktoré v jednom súvislom stringu obsahuje všetky informácie o danom produkte, ako sú hodnoty a názvy atribútov, ich preklady v inom jazyku, synonymá, ...
- Z Javy je poslaný dopyt do Elasticsearch, tomu ne nastavaný parameter fuzzy, aby Elasticsearch prípadne odhalil preklepy
- Elasticsearch vracia množinu vyhovujúcich dokumentov
- V Jave sa spracuje výsledok z Elasticsearch, zistíme, koľkokrát a v ktorých doménach a atribútoch sa nachádzajú zhody
- Na základe počítania zhôd sa vytvára strom popisujúci výsledky vyhľadávania
- Po spustení vyhľadávania sú zobrazené vyhovujúce výsledky používateľovi a po kliknutí v strome sú prefiltrované na základe domén, alebo domén a atribútov
- Grafické rozhranie je vytvorené v technológii Wicket

Analýza

- Pre analýze fulltextového vyhľadávania musíme určiť aká je úplnosť a presnosť tohto vyhľadávania
- Úplnosť je pomer vrátených relevantných výsledkov a všetkých relevantných výsledkov v indexe
- Presnosť je pomer relevantných vrátených výsledkov a všetkých vrátených výsledkov

Testovacie dopyty

	Dopyt
1	Samsung
2	Samsung 7"
3	WiFi áno Bluetooth áno
4	Kapacita baterie 2000mAh
5	Samsung Lenovo
6	Lenovo size 7
7	Sony bluetooth áno
8	Farba biela
9	Slúchadlá jack
10	Lenobo niee widi

Poďakovanie

Ďakujem RNDr. Petrovi Gurskému, PhD., vedúcemu mojej bakalárskej práce za rady a pomoc pri jej tvorbe.