

Predikcia kvality obsahu v systémoch pre odpovedanie na otázky v komunitách

Martin Borák

Vedúci práce: Ing. Ivan Srba

Jún 2015

CQA systémy

- Yahoo! Answers, Stack Exchange, Askalot
- Používateľská interakcia prostredníctvom otázok a odpovedí
- Používateľmi generovaný obsah
- Problematika kvality obsahu
 - Nízka kvalita znižuje používateľský zážitok
 - Vysoká kvalita pozitívne ovplyvňuje rast komunity
 - Kvalita obsahu napomáha kvalitnému vyhľadávaniu a odporúčaniu

Dimenzie hodnotenia kvality

- Typ obsahu
 - Otázky
 - Odpovede
- Spôsob hodnotenia
 - Analýza
 - Predikcia
- Faktory použité pri hodnotení
 - Dostupné pri pridání príspevku
 - Dostupné s odstupom času
- Výstup hodnotenia
 - Zvolenie príspevku s najvyššou kvalitou (najlepšia odpoveď)
 - Zoradenie príspevkov podľa kvality
 - Ohodnotenie kvality konkrétnou hodnotou

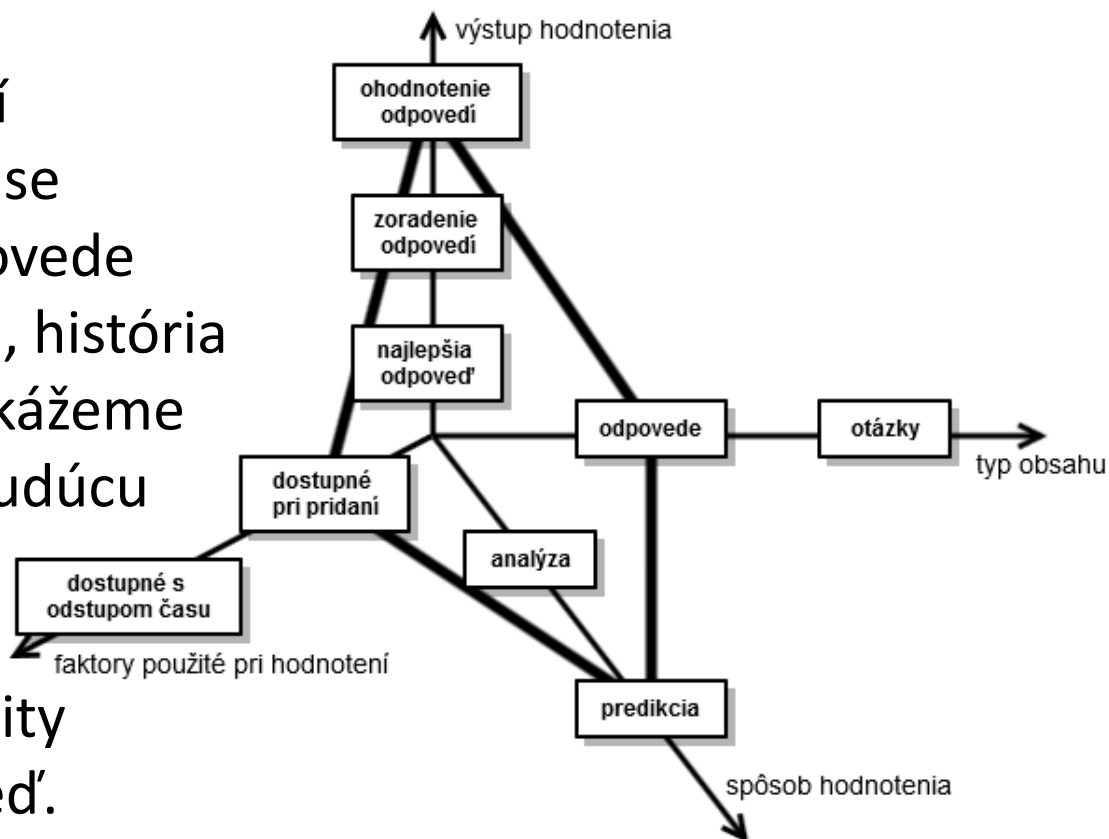
Dostupné riešenia

- Analýza kvality, do úvahy sa berie reakcia komunity → malá výpovedná hodnota pre používateľa
- Pri predikcii len jednoduchá klasifikácia
 - Predikcia najlepšej odpovede, spokojnosti používateľa
 - Využívanie klasifikačných metód
 - SVM
 - Logistická regresia

Naše riešenie

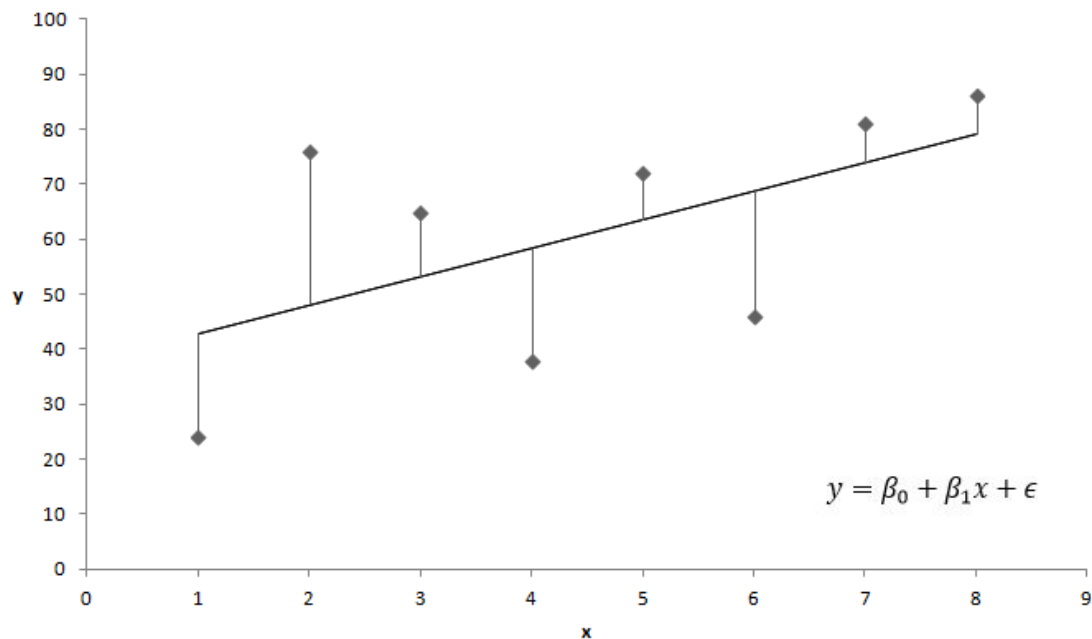
Hypotéza:

Podľa informácií dostupných v čase vytvorenia odpovede (vlastnosti textu, história používateľa) dokážeme predikovať jej budúcu kvalitu, vyjadrenú reakciou komunity na danú odpoveď.



Predikcia kvality odpovedí

- Lineárna regresia
- Nezávislé premenné – črty odpovede
- Závislá premenná – počet hlasov odpovede (za prvý mesiac od pridania odpovede)



Črty odpovedí

- Spolu 27 črt
- Črty autora odpovede (6)
 - počet otázok, odpovedí, najlepších odpovedí, čas od registrácie...
- Črty autora otázky (7)
 - priemerný počet hlasov pridelených odpovediam, ktoré reagovali na používateľove otázky...
- Črty odpovede (8)
 - počet unikátnych slov, odkazov, priemerná dĺžka slova...
- Črty otázky (4)
 - počet slov, značiek (ang. tags)...
- Ďalšie črty (2)
 - počet odpovedí, ktoré boli pridane k otázke pred vytvorením vyhodnocovanej odpovede, priemerný počet hlasov k týmto odpovediam

Extrakcia textových črt

- Odstránenie diakritiky
- Lematizácia
- Odstránenie stopslov

Aký je význam slova "léma"?



Aký je význam slova léma



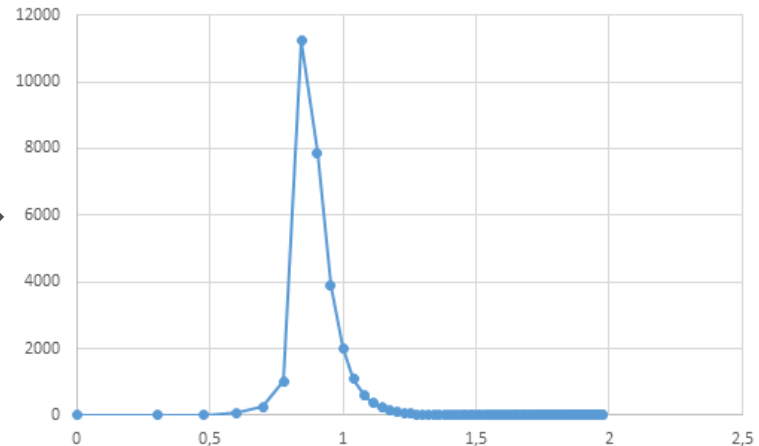
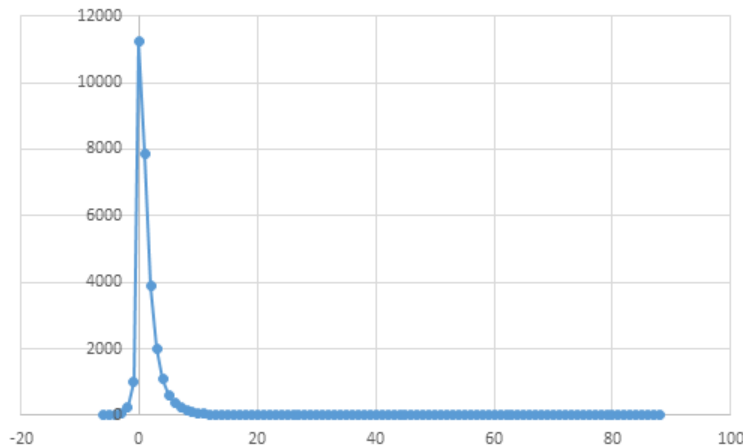
Aký byť význam slovo léma



význam slovo léma

Úprava hodnôt po extrakcii

- Normalizácia hodnôt extrahovaných črt na interval $<0; 1>$
- Zlogaritmovanie počtu hlasov → snaha o odstránenie problému distribúcie typu „dlhý chvost“ (ang. long tail)



Implementácia a experimentálne overenie

- Implementácia
 - Experimentálna infraštruktúra systému Askalot
 - Ruby on Rails
- Overenie
 - Dátová sada zo systému Android z rodiny Stack Exchange (takmer 30 000 odpovedí z rozmedzia 5 rokov)
 - Využitá K-násobná krížová validácia

Výsledky – štatistiky

Štatistický údaj	Hodnota
R	0,388
R ²	0,151
R ² Adj	0,150
Štandardná odchýlka	0,083

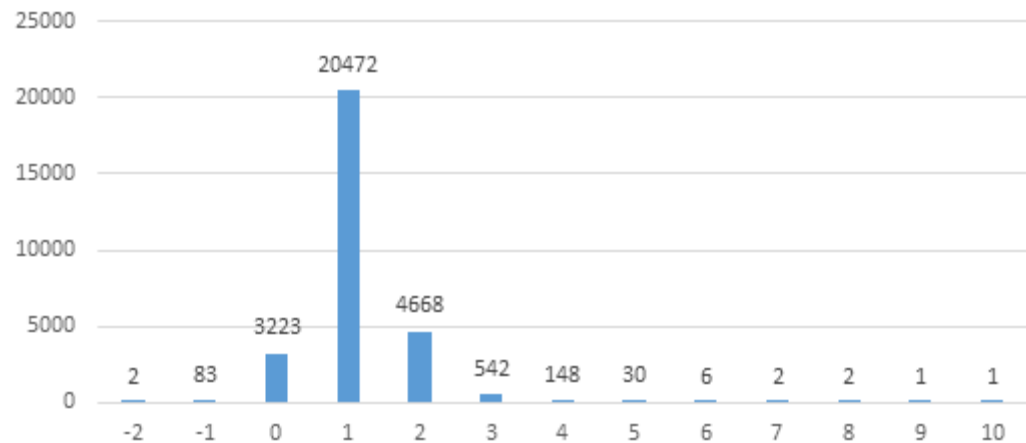
- R – korelácia medzi nezávislými premennými a závislou premennou
- R² – kovariancia, určuje aká časť pohybu nezávislej premennej je spôsobená závislými premennými
- R² Adj – ako predchádzajúca, no berie do úvahy početnosť vzorky, na ktorej regresia prebehla

Výsledky

Typ vyhodnocovania	Počet správnych (z celkových 28 920)	Presnosť
presné	8 398	29,00 %
odchýlka ± 1	22 082	76,36 %
$\pm$ štatistická odchýlka	22 265	76,99 %

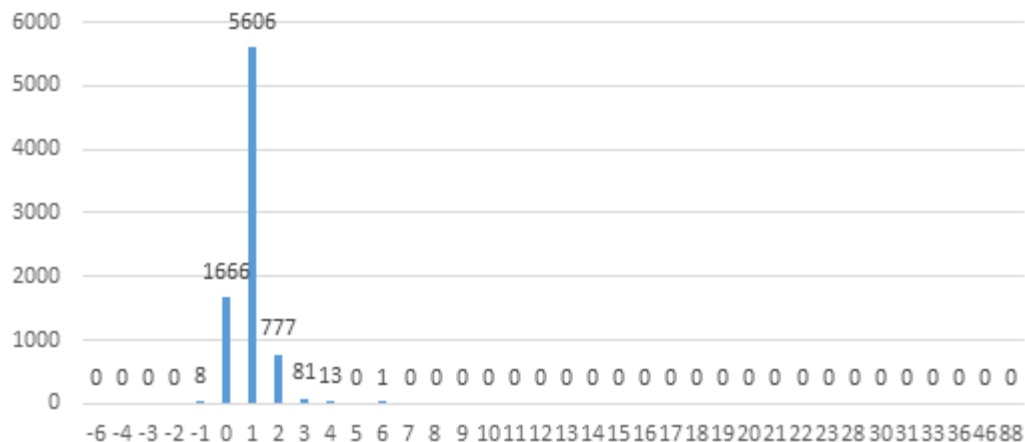
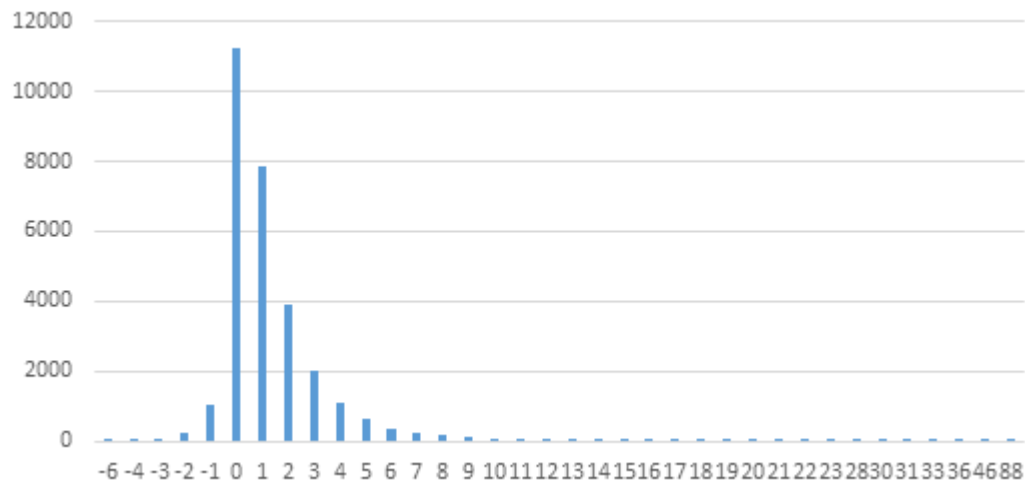
- Presné – správne odhadnutý konkrétny počet hlasov
- Odchýlka ± 1 – počet hlasov s prípadnou chybou jedného hlasu
- $\pm$ štatistická odchýlka – počet hlasov s prípadnou chybou určenou štatistickou časťou výstupu lineárnej regresie

Histogram reálnych hodnôt



Histogram predikovaných hodnôt

Histogram správne predikovaných hodnôt



Črty s najväčšou váhou

1. Priemerný počet hlasov pridelených predchádzajúcim odpovediam autora odpovede
2. Počet unikátnych slov v odpovedi
3. Priemerný počet hlasov k predchádzajúcim odpovediam
4. Počet najlepších odpovedí, ktoré autor odpovede pridal
5. Počet odkazov v odpovedi

Zhrnutie

- Metóda na predikciu počtu hlasov udelených odpovedi v CQA systémoch hneď po jej pridaní
- Extrakcia textových črt a črt z profilov používateľov
- Využitie lineárnej regresie
- Úspešnosť (29%) obmedzená nerovnomernou distribúciou počtu hlasov v CQA systémoch a nedostatočným predikčným potenciálom črt dostupných pri pridaní odpovede