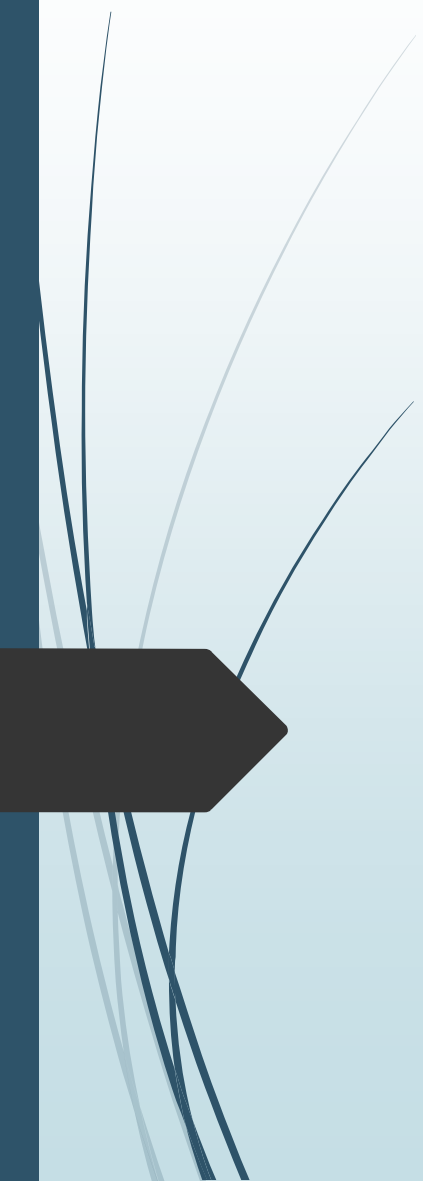




Predikcia popularity vedeckých článkov

autor: Adam Bacho | vedúci: Ing. Róbert Móro



Motivácia

- množstvo nových článkov vychádzajúcich každý deň/týždeň/mesiac
 - rôzne konferencie, vedecké časopisy, autori
- počet citácií, ktoré článok získa, je jednou z možných meradiel kvality týchto článkov
 - problém zistiť kvalitu článkov, ktoré boli vydané len nedávno



Cieľ

- dokázať predpovedať počet citácií článkov N rokov po ich vydaní
 - toto má byť dosiahnuté čo najskôr po ich vydaní
 - zahrnutie črt, ktoré ešte pri predikcii použité neboli – napr. Eigenfactor



Stav v oblasti

- rôzne výsledky pre rovnaké črty
- najčastejšou metódou strojového učenia je lineárna regresia
 - niektorí autori sa však zamerali aj na porovnanie vplyvu rôznych metód
- najúspešnejšou črtou sa ukázal byť Journal Impact Factor
 - nie všetci autori ho však rešpektujú
- veľká časť autorov volí peer-review metódu a pracuje s relatívne malým počtom článkov



Návrh metódy predikcie počtu citácií – ČRTY

- **črty článkov** – počet strán, autorov, inštitúcií, kľúčových slov, referencií, vývoj počtu citácií v čase, atď.
- **črty vedeckých časopisov** – Journal Impact Factor, Eigenfactor
- **črty alternatívnych metrík** – počet zdieľaní na sociálnych sieťach (Facebook, Twitter, Google+, atď.)
- **rozdiel črty Eigenfactor od JIF**
 - referencie článkov v rámci jedného vedeckého časopisu sú odstránené – zaniká sebacitovanosť
 - citácie ovážené podľa kvality vedeckého časopisu

Návrh metódy predikcie počtu citácií – METÓDA STROJOVÉHO UČENIA

- **lineárna regresia** – cieľom nebolo porovnávať metódy strojového učenia, preto bola vybratá lineárna regresia, ktorá je používaná najčastejšie

$$\sum_{i=1}^n \left(x^i - \sum_{j=0}^k w_j a_j^{(i)} \right)^2$$



Návrh metódy predikcie počtu citácií – DÁTOVÁ SADA

- **PubMed** – viac ako 900 000 článkov z viac ako 4 700 vedeckých časopisov
 - najdôležitejším kritériom, ktoré rozhodlo o jej výbere bol jednoznačný identifikátor článkov PubMed ID a ISSN
 - články v XML formáte – potreba predspracovania
 - nie všetky vybrané črty obsiahnuté v dátovej sade – potreba ich získania z ISI Web of Knowledge
- po doplnení požadovaných črt z externých zdrojov ostalo z pôvodných článkov len niekoľko tisíc v každom zo skúmaných rokov

Experimentálne overenie

- bolo vytvorených niekoľko modelov na základe rôznych črt
 - črty článkov, informácie o vývoji počtu citácií v čase, JIF, Eigenfactor
- každý model sme otestovali pomocou 10-krát opakovanej 10-zložkovej krížovej validácií
- merítka úspešnosti boli
 - R^2
 - RMSE (root mean squared error)

Naše výsledky 1 – model na základe štatisticky významných črt

- do tohto modelu vstupovali články z roku 2005 a bol predpovedaný počet citácií 8 rokov po vydaní článku

NÁZOV ČRTY	REGRES. KOEF.	P-HODNOTA	R ² (%)	RMSE
počet znakov v nadpise	-6,6	< 0,05	8,24	10,1
počet slov v nadpise	1,48	0,54		
počet znakov v abstrakte	8,7	0,17		
počet slov v abstrakte	-7,46	0,25		
počet referencií	31,09	< 0,001		
počet strán	55,79	< 0,001		
počet autorov	19,63	< 0,001		
počet inštitúcií	17,08	< 0,001		
počet kľúčových slov	-7,95	< 0,001		

Naše výsledky 2 – model po pridání vývoja počtu citácií v čase

- predchádzajúci model bol obohatený o informáciu o vývoji počtu citácií v čase

MODEL PO PRIDANÍ CITÁCIÍ ZÍSKANÝCH:	R² (%)	RMSE
v roku vydania článku	17,8	9,54
do 1 roka po vydaní článku	42,3	8,1
do 2 rokov po vydaní článku	58,3	6,93
do 5 rokov po vydaní článku	89,8	3,45

Naše výsledky 3 – model pre porovnanie črt JIF a Eigenfactor

- na základe dostupnosti črt JIF a Eigenfactor bol tento model postavený na článkoch z roku 2010 a bol predpovedaný počet citácií 3 roky po vydaní článku

ČRTY	R ² (%)	RMSE
JIF	17,8	8,02
Eigenfactor	10,3	8,34
JIF & Eigenfactor	17,6	8,04



Zhrnutie



Dosiahnuté výsledky

- JIF silnejším prediktorom ako Eigenfactor
- vysoký vplyv vývoja počtu citácií v čase na ich konečný počet
- štatisticky významnými prediktormi sa ukázali byť počet referencií, strán, autorov a počet inštitúcií

Priestor na vylepšenia

- tvorba vlastnej dátovej sady alternatívnych metrík
- porovnanie výsledkov pre rôzne metódy strojového učenia
- aplikácia modelu medzi doménami