Personalizované odporúčanie s využitím kontextu pre nových používateľov

Diplomová práca

Bc. Róbert Kocian
Vedúci: Ing. Michal Kompan, PhD. 11.6.2015

Motivácia a problém

- Nový používateľ - dôležitý prvý dojem
- Kľúčové relevantné odporúčanie
- Problém studeného štartu - nedostatok informácii o novom používateľovi

Používané metódy odporúčania pre problém studeného štartu

- Faktorizácia matíc[1]
  - \+ Redukuje problém „riedkej“ matice rozdelením matíc
  - \+ Dobre reaguje na nové položky
  - \- Ťažšie hľadanie skrytého faktora
- Kombinované odporúčanie (obsahové a kolaboratívne)[1]
  - \+ Zložitosť
  - \+ Redukuje problém „riedkej“ matice prehľadávaním textu
  - \- Menej pružná na pridávanie nových položiek

---

[1]Adomavicius 2011

Doména digitálnych knižníc

- Pravidelný cyklický prísun nových používateľov – ročný interval
- Denne pribúda malo článkov

- Problémy v oblasti digitálnych knižníc:
  - Spracovanie veľkého množstva dokumentov a ich textu
  - Viacznačnosť slov
  - Dokumenty môžu byt v rôznych jazykoch

Kontext nového používateľa

- Pod kontextom používateľa rozumieme dáta:
  - Informačný systém AIS
    - Záverečné práce
    - Rok a typ záverečnej práce
    - Práce študentov vedúceho
  - Publikačný portál STU
    - Publikácie vedúcich

Riešenie

- Zohľadnením kontextu nového výskumníka môžeme zlepšiť presnosť generovaného personalizovaného odporúčania v počiatočných fázach interakcie zo systémom.

Návrh metódy odporúčania

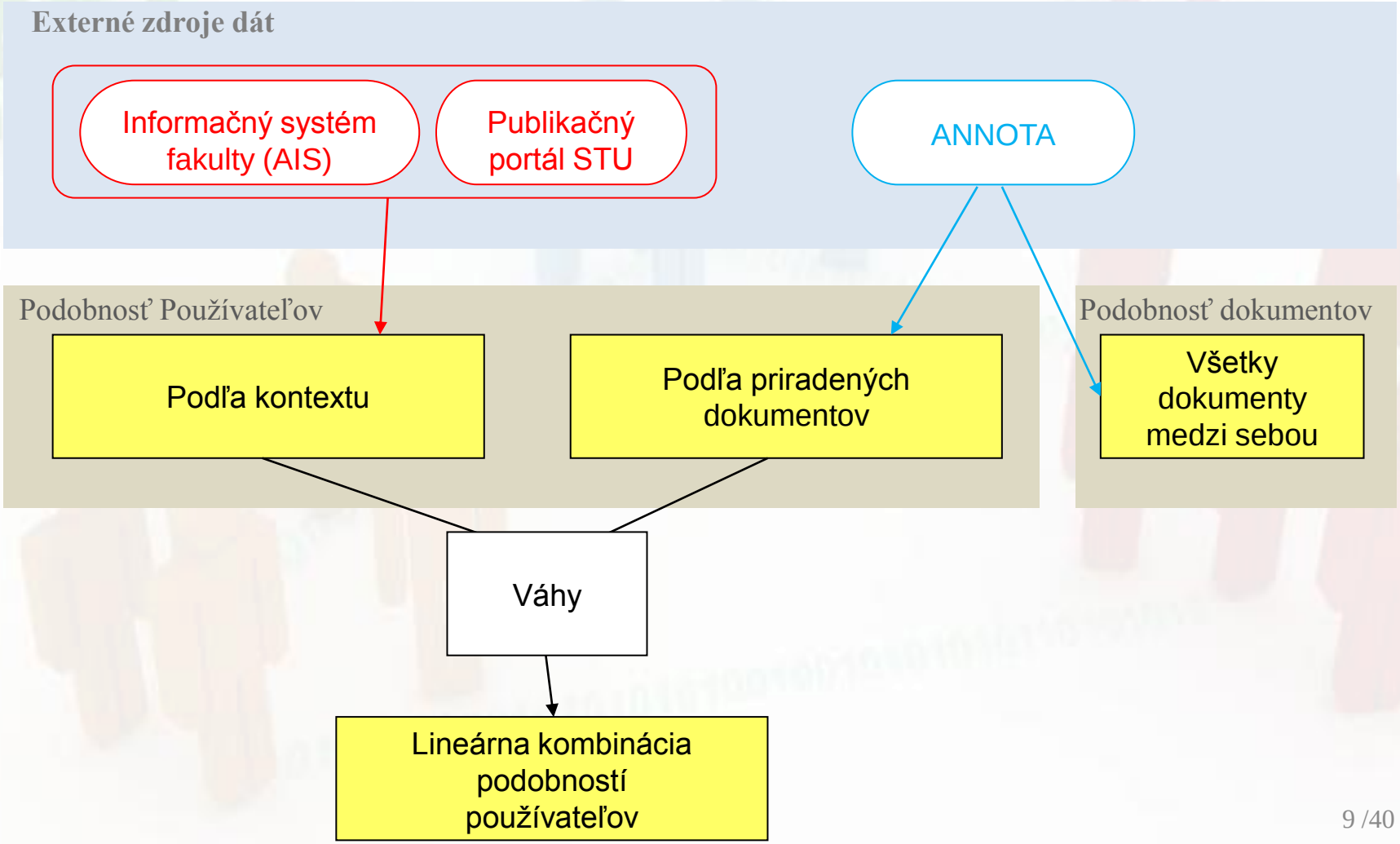
- Metóda sa skladá sa z dvoch základných častí
  - Krok 1. Metóda podobnosti
    1. Predpracovanie dát
    2. Výpočet podobnosti
  - Krok 2. Metóda odporúčania
    1. Výpočet skóre dokumentov
    2. Zoradenie dokumentov podľa skóre

Krok 1. Metóda podobnosti

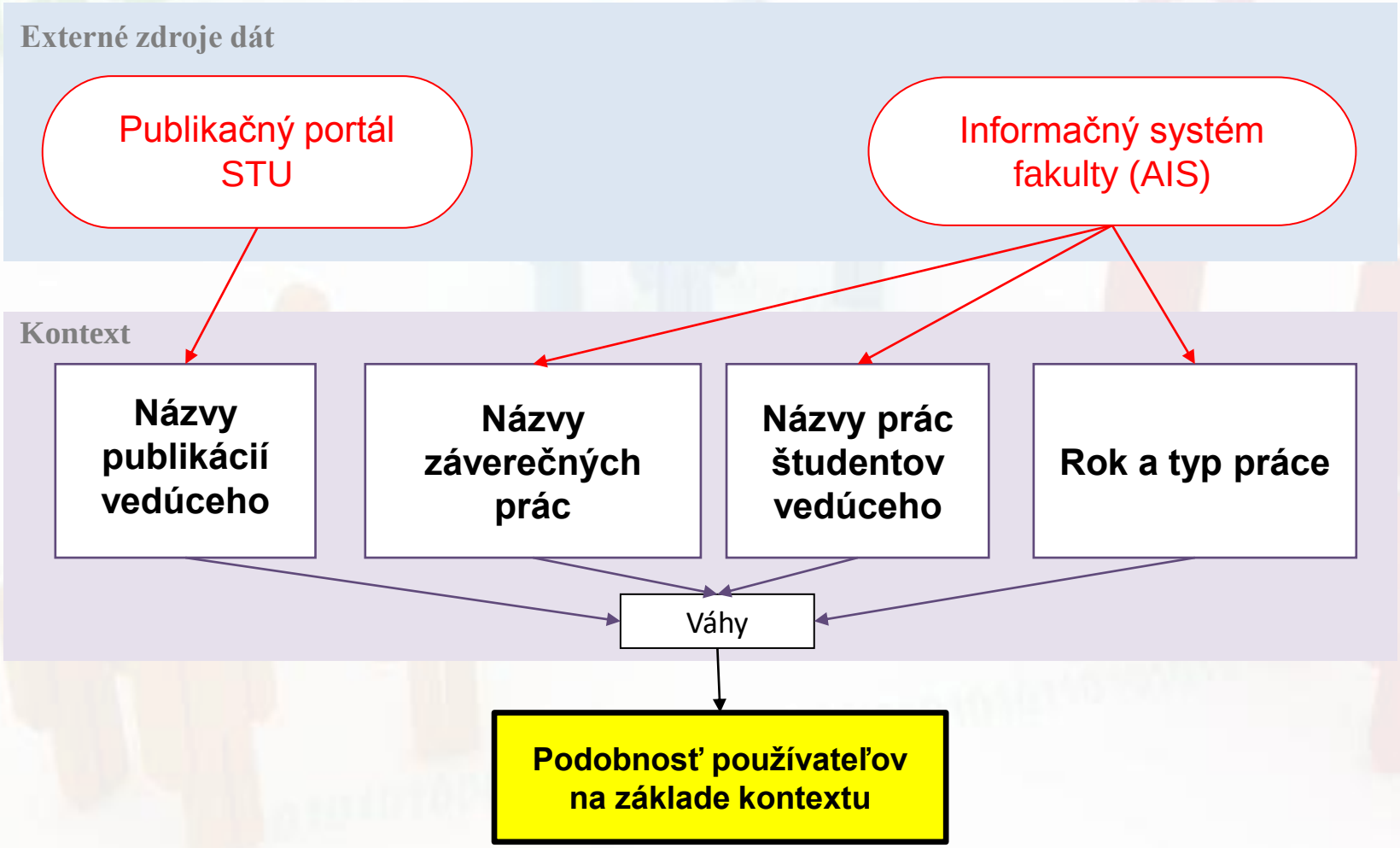
1. Predspracovanie textu:
   - Stop slová
   - Lexikálna analýza
   - Redukcia na koreň slova - stemovanie

- Váhy rôznych typov podobností

Krok 1. Metóda podobnosti



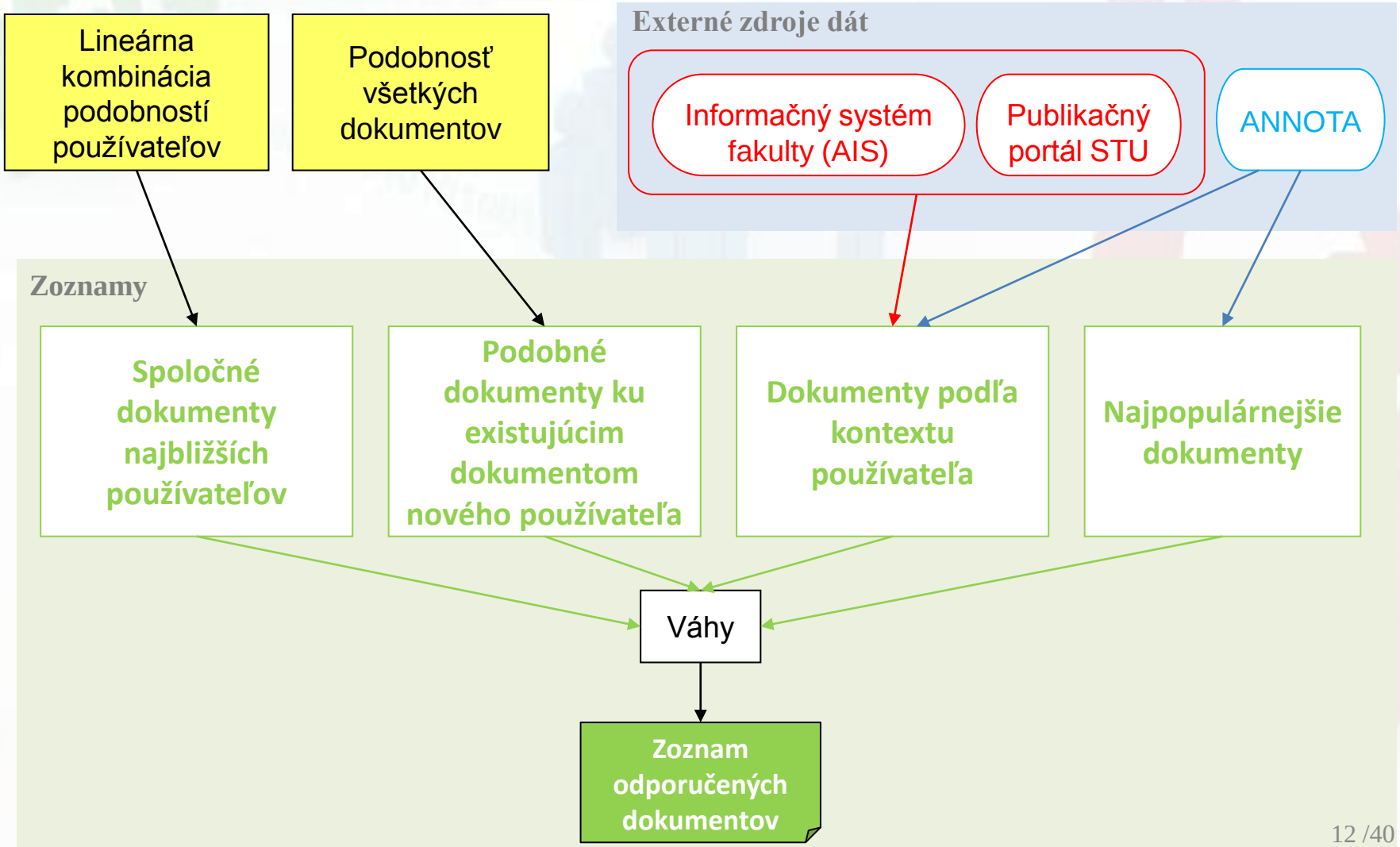
Krok 1. Metóda podobnosti



Krok 2. Metóda odporúčania

1. Výpočet skóre dokumentov
2. Zoradenie dokumentov podľa skóre

Krok 2. Metóda odporúčania



Krok 2. Metóda odporúčania – definovanie skóre a zoradenie zoznamov

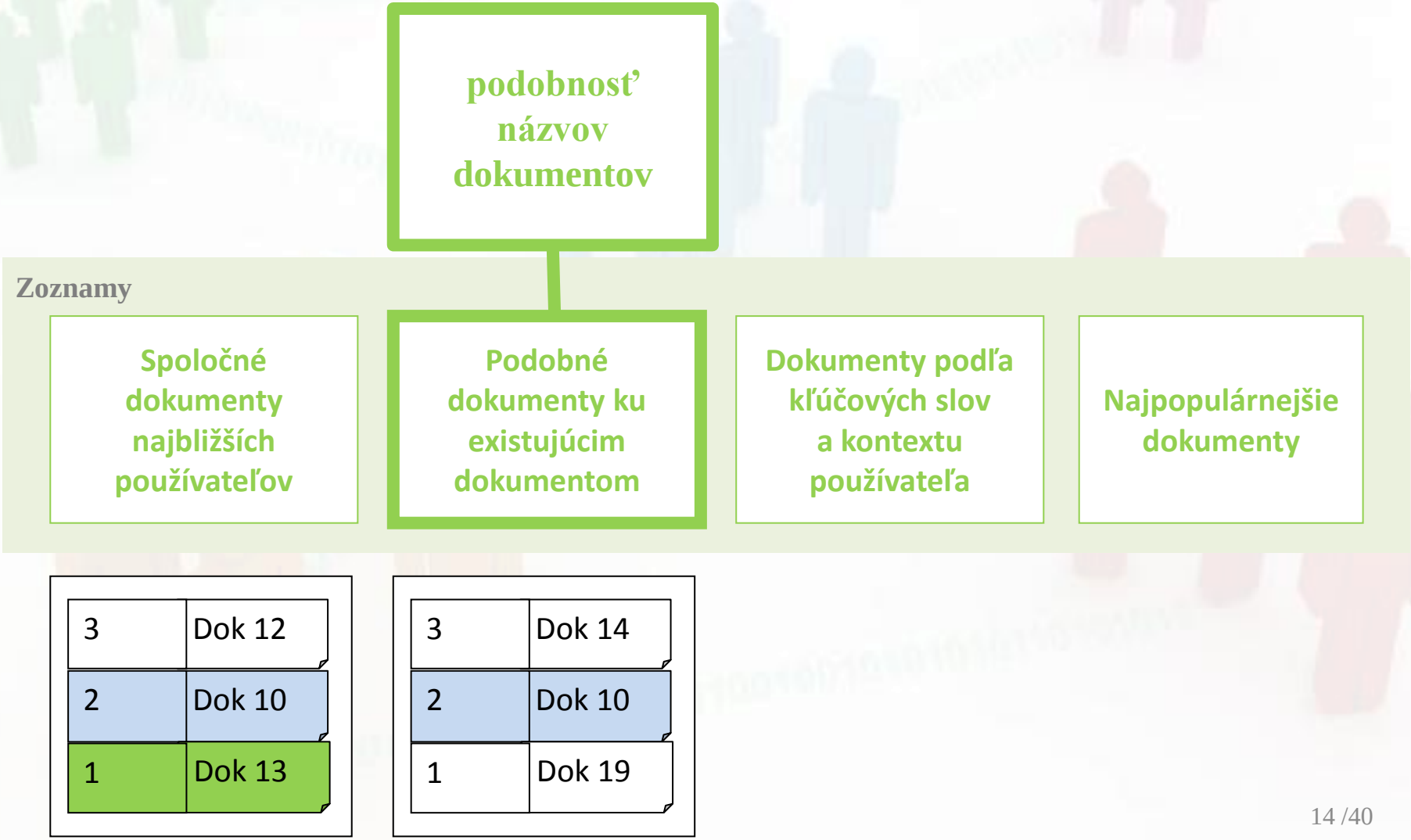
Počet spoločných dokumentov ku počtu najbližších používateľov

Zoznamy

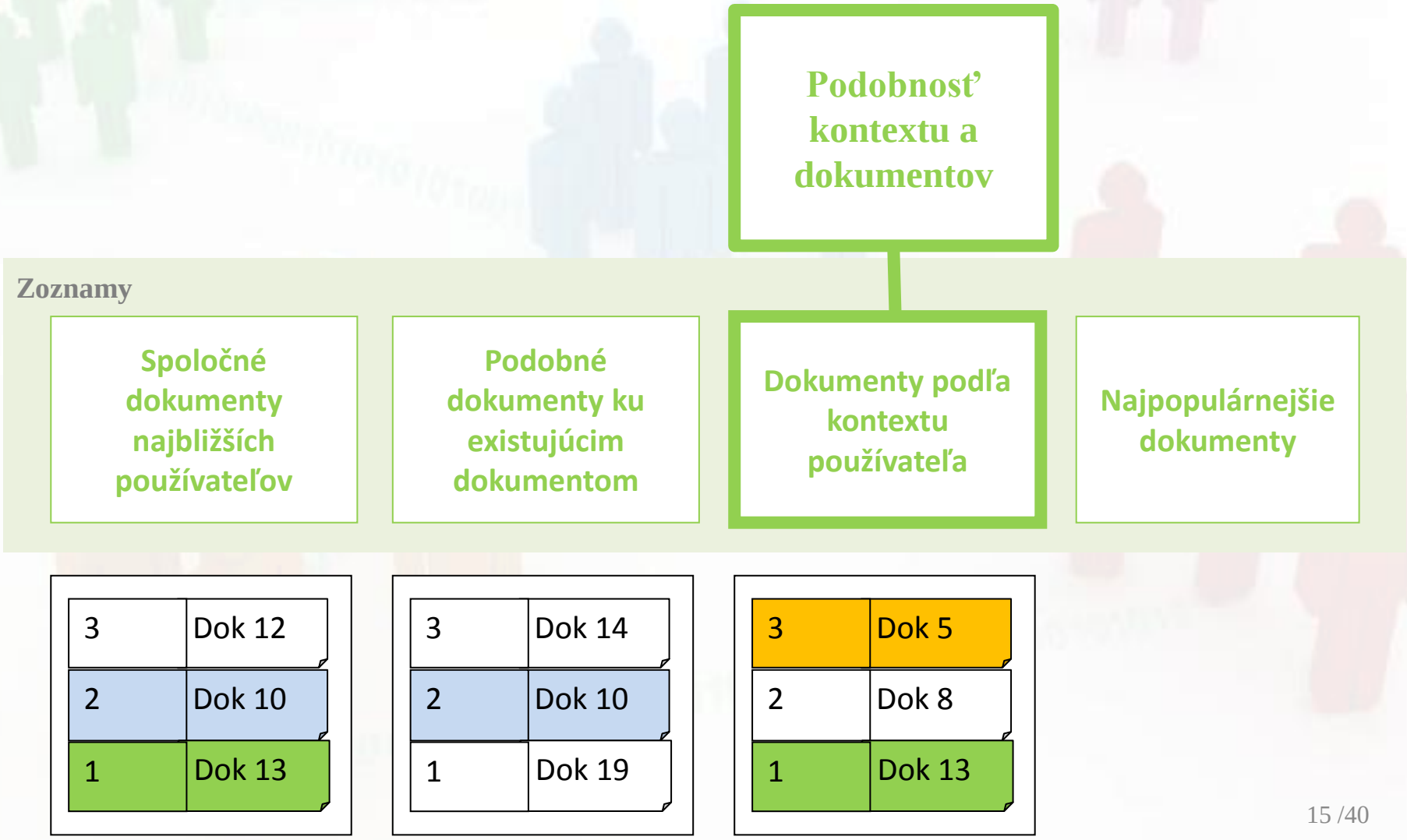
- **Spoločné dokumenty najbližších používateľov**
- **Podobné dokumenty ku existujúcim dokumentom nového používateľa**
- **Dokumenty podľa kľúčových slov a kontextu používateľa**
- **Najpopulárnejšie dokumenty**

3	Dok 12
2	Dok 10
1	Dok 13

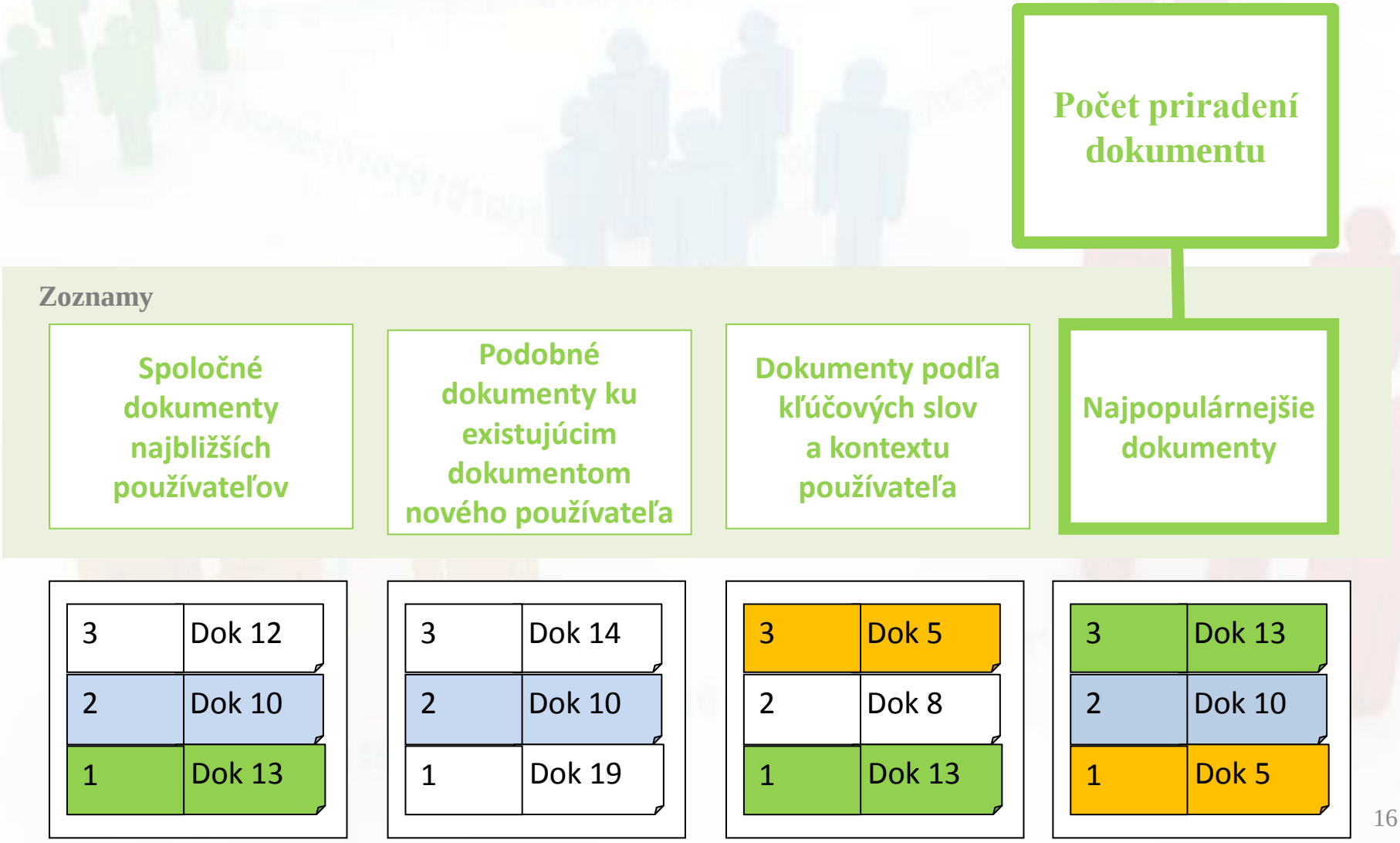
Krok 2. Metóda odporúčania – definovanie skóre a zoradenie zoznamov



Krok 2. Metóda odporúčania – definovanie skóre a zoradenie zoznamov



Krok 2. Metóda odporúčania – definovanie skóre a zoradenie zoznamov



Realizácia

- C# .Net 4.5
- Knižnica numl.net – práca s maticami a vektormi
- Databáza PostgreSQL 9.3
- Pre výpočet podobností bola použitá kosínusová podobnosť
- Metóda odporúčania bola implementovaná ako webová služba
- TDD – metodika vývoja
- Trojvrstvová architektúra:
  - Dátová vrstva
  - Aplikačná vrstva
  - Prezentačná vrstva

Overenie v doméne digitálnych knižníc

- Časové obdobie experimentov 1,5 mesiaca
- Offline experimenty s dátami:
  - ANNOTA:
    - už existujúci používatelia ANNOTA aplikácie
    - 280 používateľov v aplikácii ANNOTA
  - Publikačný portál STU:
    - Počet všetkých publikácii vedúcich 227
  - AIS:
    - 997 záznamov záverečných prác v časovom horizonte od 2010 až 2014

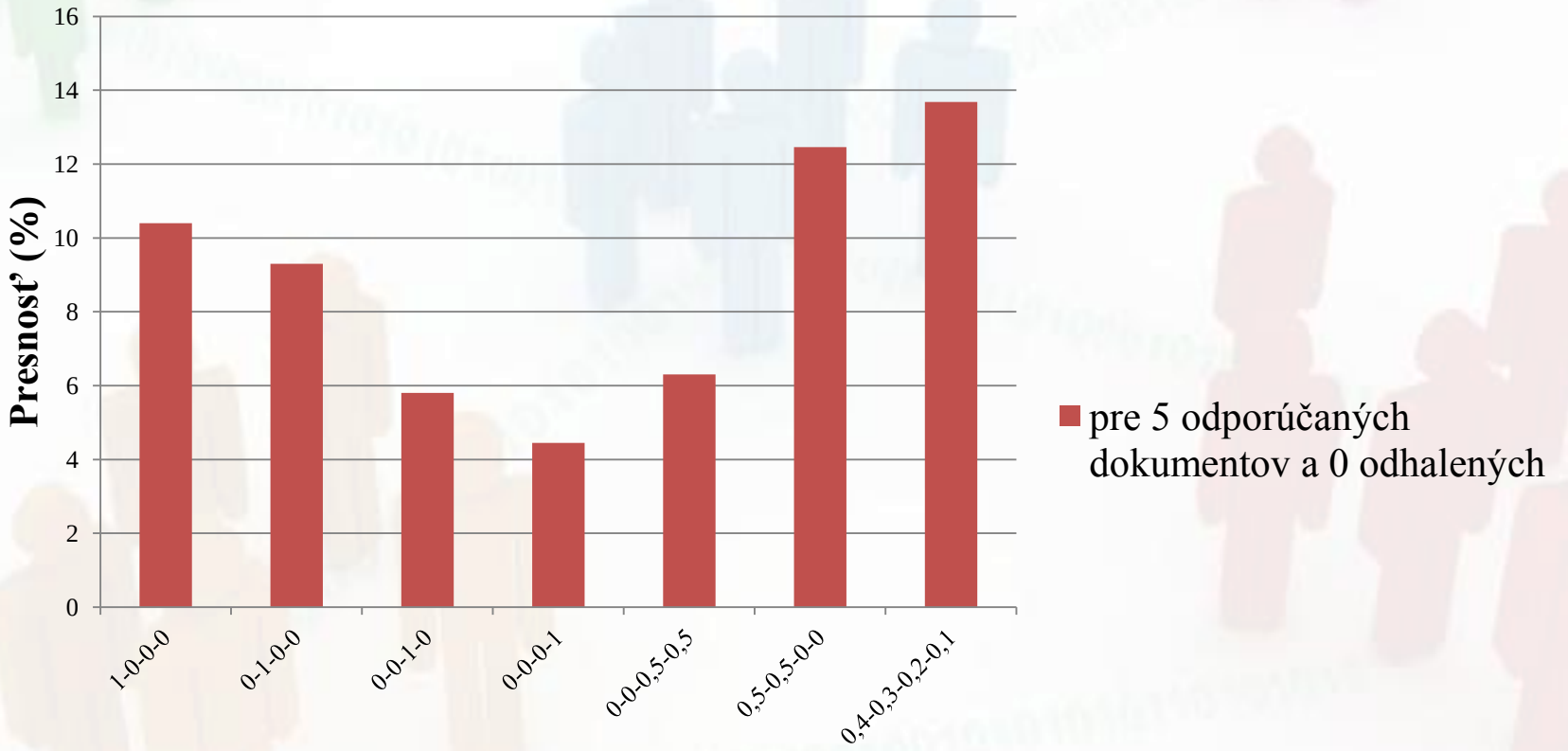
Overenie v doméne digitálnych knižníc - postup

1. Vybratie všetkých používateľov z ANNOTA, ktorí majú aspoň kontext AIS
2. Stanovenie maximálneho počtu odporúčaných dokumentov – 3,5 a 10
3. Postupné odkrývanie dokumentov od 0 až po 10 dokumentov
4. Nastavenie váh
5. Výpočet podobnosti a odporúčania
6. Výpočet presnosti
   - počet správne odporúčaných / maximálny počet odporúčaných

Hypotéza

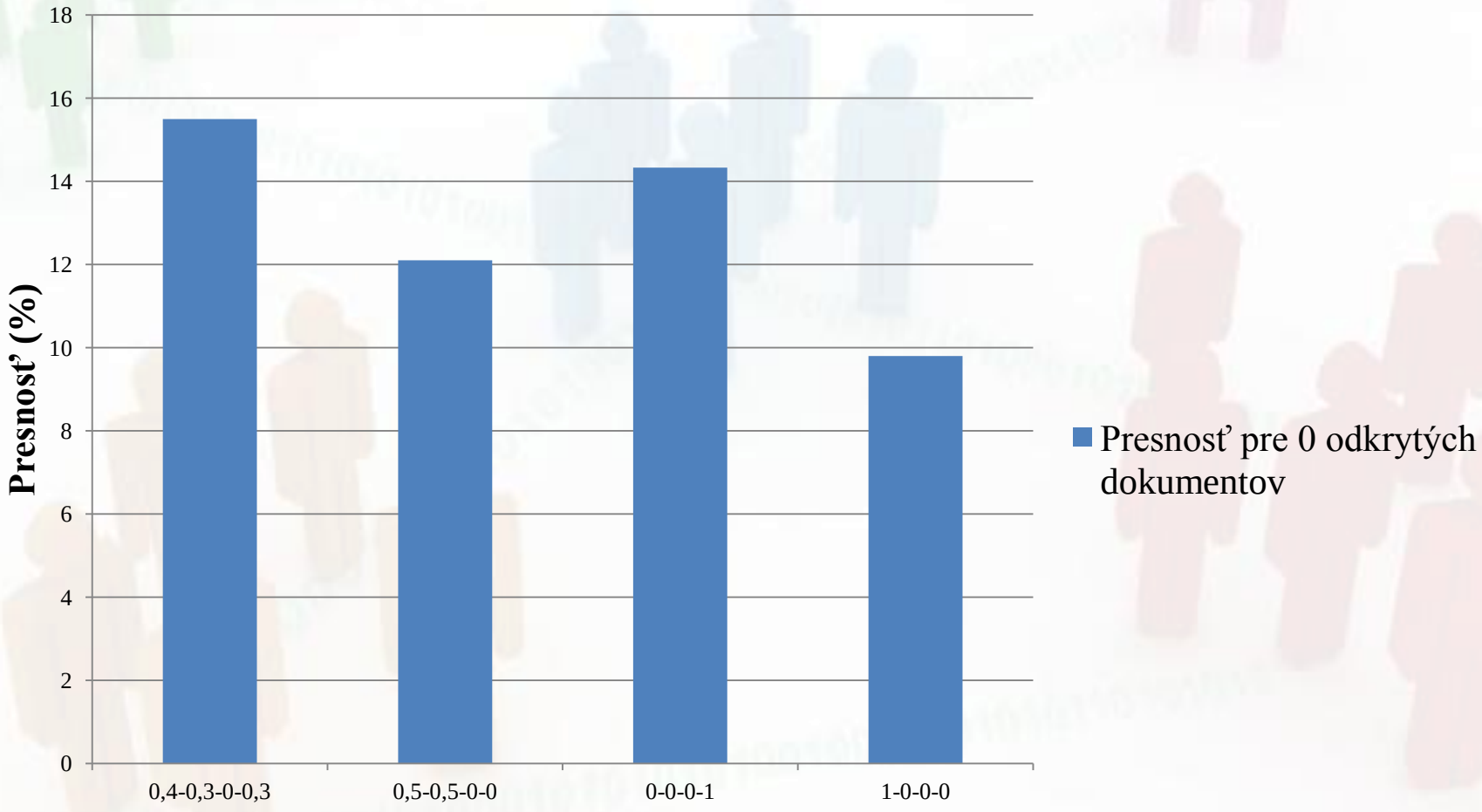
- Navrhnutá metóda odporúčania dokáže pridávaním kontextu používateľa zlepšiť presnosť odporúčania pre nového používateľa v porovnaní s metódou odporúčania najpopulárnejších článkov.

Experimentálne overenie váh kontextu pre 34 používateľov



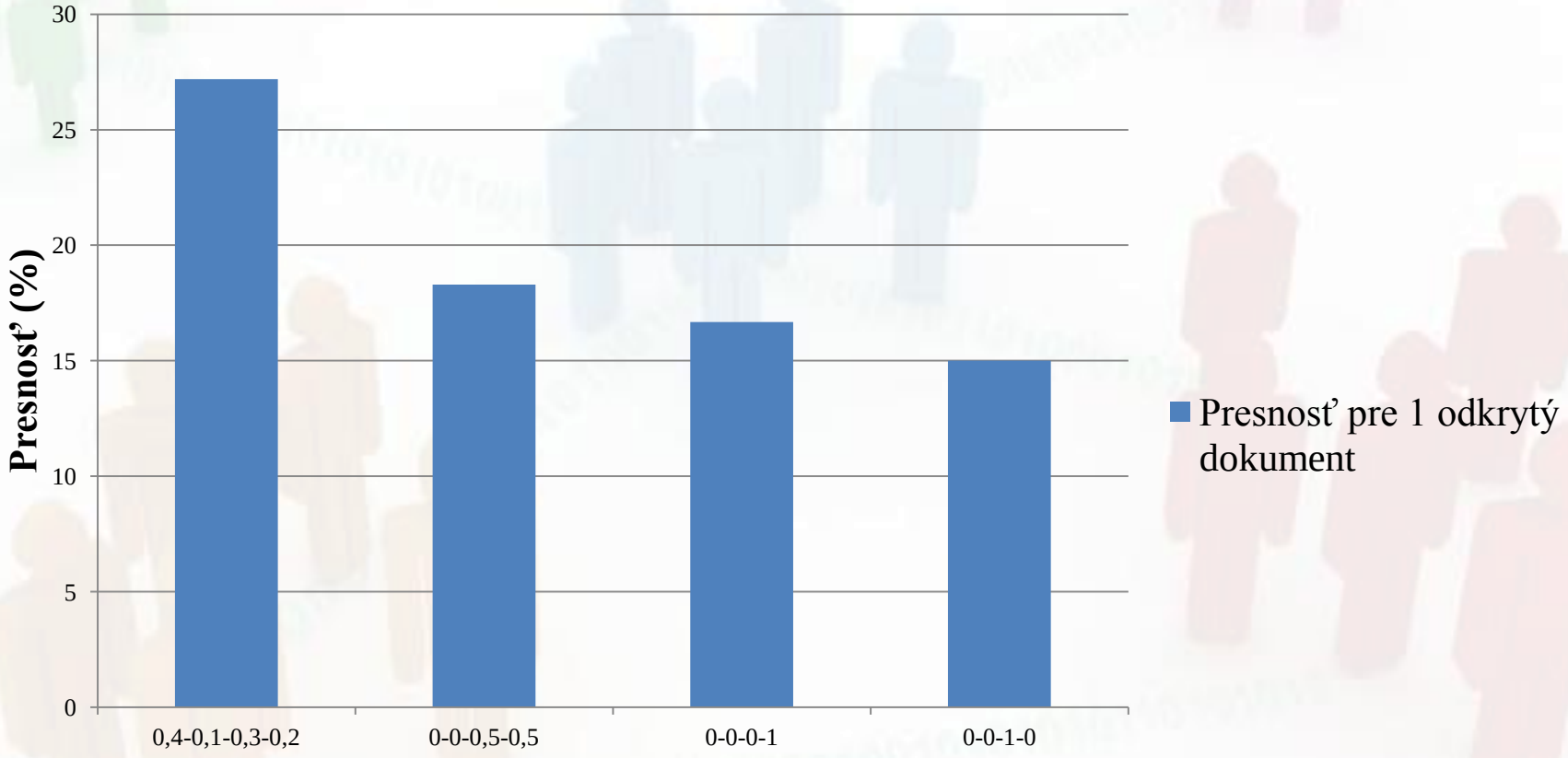
kontext : školské projekty - publikácie vedúceho - projekty študentov vedúceho - čas

Experimentálne overenie váh zoznamov pre 3 odporúčané dokumenty a 43 používateľov



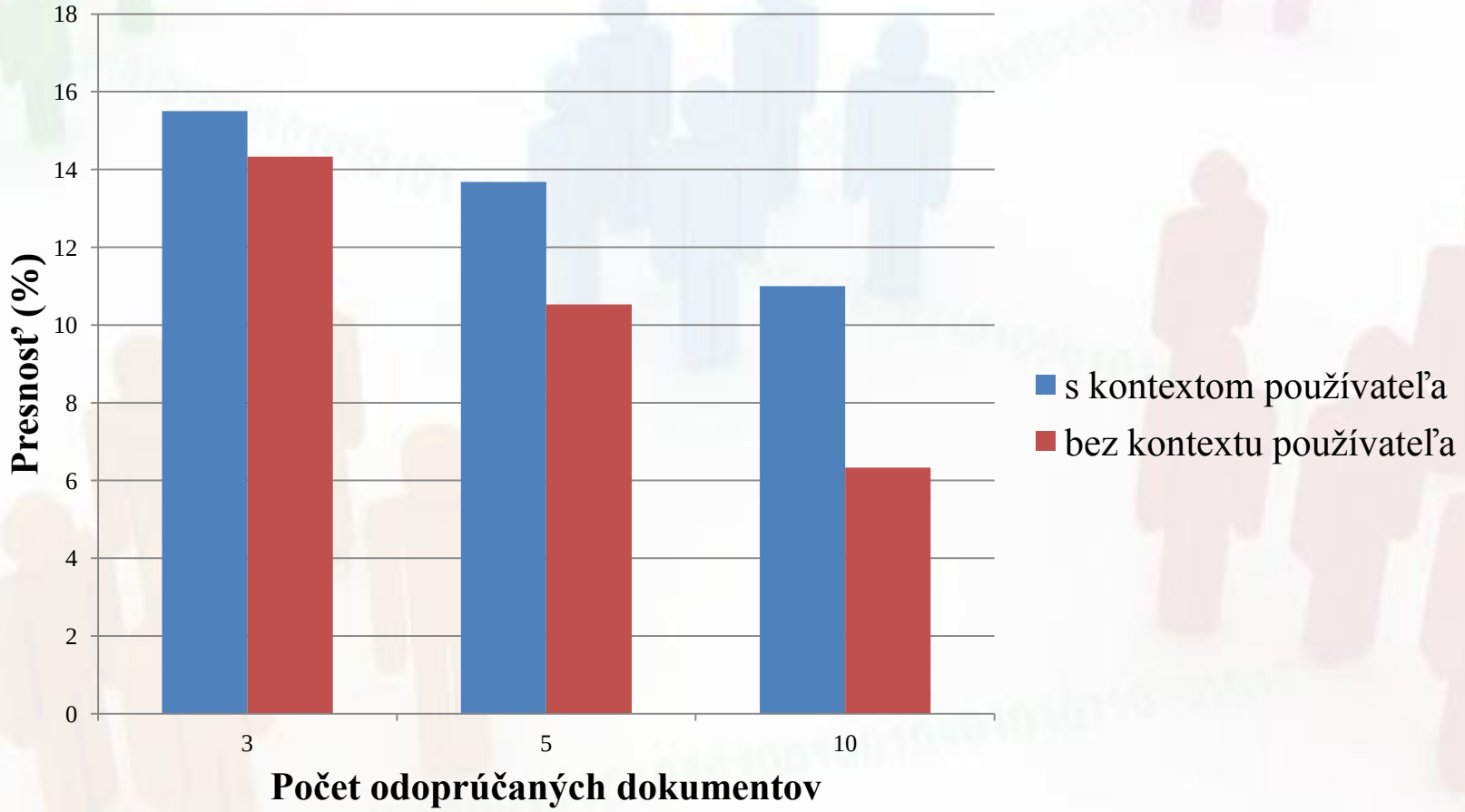
Váhy: spoločné dokumenty - kľúčové slová dokumentov - podobnosť s existujúcimi dokumentmi - najpopulárnejšie dokumenty

Experimentálne overenie váh zoznamov pre 3 odporúčané dokumenty a 38 používateľov



Váhy: spoločné dokumenty - kľúčové slová dokumentov - podobnosť s existujúcimi dokumentmi - najpopulárnejšie dokumenty

Experiment - prínos kontextu pre 0 odkrytých dokumentov



Prehľad výsledkov

Počet odporúčaných dokumentov	Počet používateľov	Počet odkrytých dokumentov	**Naša metóda**		Metóda najpopulárnejších dokumentov	
			Presnosť (%)	**Počet používateľov s nulovou presnosťou**	Presnosť (%)	Počet používateľov s nulovou presnosťou
3	43	0	**15,50**	**28**	14,33	29
5	38	0	**13,68**	**20**	10,53	24
10	30	0	**11,01**	**13**	6,33	17
3	38	1	**27,19**	**17**	16,67	25
5	34	1	**22,94**	**14**	11,18	21
10	28	1	**22,50**	**7**	6,43	16
3	35	4	**40,00**	**10**	18,09	22
5	34	4	**38,24**	**7**	11,18	21
10	28	4	**30,71**	**5**	6,43	16
3	28	10	**42,48**	**7**	21,43	16
5	28	10	**39,98**	**5**	11,18	16
10	28	10	**35,43**	**5**	6,43	16

Zhodnotenie

- Prínosom práce je navrhnutá metóda odporúčania, ktorá:
  - Odporúča aj pre existujúceho používateľa
  - Dynamické pridávanie kontextu
- Zlepšenie v priemere o 12% už pri prvom dokumente používateľa
- Možné použitie pre rôzne typy aplikácií v doméne digitálnych knižníc
- Možnosť vylepšenia metódy - sémantika slov

Používateľská prezentácia

Živá ukážka

Používateľská prezentácia

- Navrhnutá metóda bola implementovaná ako webová služba poskytujúca odporúčané dokumenty
- Vlastnosti:
  - Časová optimalizácia počítania podobností
  - Asynchrónna komunikácia
  - Jednoduchosť implementácie na klientskej strane pomocou AJAX

Vstup

- Vstupom je identifikačné číslo prihláseného používateľa

- Príklad vstupu pomocou GET:
  - http://147.175.146.184/recservice/ColdStartService.svc/getrec/275

Výstup

- Výstupom je JSON objekt obsahujúci:
  - ID dokumentu
  - Názov dokumentu
  - Odkaz na detail dokumentu

Výstup

HTTP/1.1 200 OK
Cache-Control: private
Content-Length: 483
Content-Type: application/json; charset=utf-8
Server: Microsoft-IIS/7.5
X-AspNet-Version: 4.0.30319
X-Powered-By: ASP.NET
Date: Mon, 08 Jun 2015 10:48:54 GMT
[
{"DocumentID":1,"DocumentLink":"http:\/\/link.springer.com\/chapter\/10.1007\/978-0-387-85820-3_7","DocumentTitle":"Context-Aware Recommender Systems - Springer"},

{"DocumentID":2,"DocumentLink":"http:\/\/dl.acm.org\/citation.cfm?id=996402","DocumentTitle":"A Hybrid Tag-Based Recommendation Mechanism to Support Prior Knowledge Construction"},

{"DocumentID":2,"DocumentLink":"http:\/\/dl.acm.org\/citation.cfm?id=996402","DocumentTitle":"Enhancing digital libraries with TechLens+"}
]

Záver

- Silné stránky
  - Jednoduchá implementácia
  - Časová optimalizácia výpočtov
  - Asynchrónna komunikácia
- Slabé stránky
  - Prenositeľnosť na iné aplikácie

Zhodnotenie

- Prínosom práce je navrhnutá metóda odporúčania, ktorá:
  - Odporúča aj pre existujúceho používateľa
  - Dynamické pridávanie kontextu
- Zlepšenie v priemere o 12% už pri prvom dokumente používateľa
- Možné použitie pre rôzne typy aplikácií v doméne digitálnych knižníc
- Možnosť vylepšenia metódy - sémantika slov

Posudok oponenta - vyjadrenie

- Časť overenie (hypotéza)
  - Primárnym cieľom navrhnutej metódy bolo overiť či kontext môže zlepšiť výsledok odporúčania.
  - Ako môžeme pozorovať v experimente s hľadaním vhodných váh tak, pri nesprávnom použití váh môže byť výsledok presnosti menší aj keď máme zahrnuté najčítanejšie dokumenty, čiže metóda nebude vždy s kontextom úspešná.
  - Je to z dôvodu niektorých špecifických názvov záverečných prác, ktoré nie sú porovnateľné s inými

Posudok oponenta – otázka 1

- V overení ste našli váhy vstupných signálov 0,4; 0,3; 0,2; 0,1 pre ktoré bola presnosť odporúčania najlepšia, akým algoritmom ste tieto váhy našli?
- Nakoľko sú tieto váhy vhodné pre použitie na iných (nových) dátach?
- Poznáte metódy ako vyhodnotiť vhodnosť váh pre použitie s inými dátami?

Posudok oponenta – odpoveď 1

- Na hľadanie najlepších váh som použil postupné zvyšovanie hodnôt pre každú váhu a kombinácie medzi nimi.
- Ak by iné dáta boli podobného typu kontextu s najväčšou pravdepodobnosťou by ich v tomto nastavení bolo možné použiť.
- Vhodnosť váh pre iné dáta by som vyhodnotil rovnakou metódou ale je možné použiť aj viacrozmerné optimalizačné metódy hľadania extrému funkcie.

Posudok oponenta – otázka 2

- Overenie prispôsobovania odporúčania histórií dokumentov je vyhodnotené v práci obmedzene. Ako sa mení presnosť odporúčania s pribúdajúcimi (odkrytými) dokumentmi? Ako sa menia „najlepšie“ váhy?
- Treba ich prispôsobovať?
- Ak áno, je to zovšeobecniteľné?

Posudok oponenta – odpoveď 2

- Pri viac ako 10 odkrytých dokumentov sa zvyšuje váha podobnosti s existujúcimi dokumentmi a od 50 dokumentov je presnosť medzi 50-60%
- Je potrebné ich prispôsobovať pretože pri rastúcom počte odhalených dokumentov je presnosť vyššia ak je väčšia váha pre podobnosť s už existujúcimi dokumentmi
- Áno je to zovšeobecniteľné pretože kontext môže pomôcť pri počiatočnej fáze, kedy je riedkosť matice veľká ale ak máme dáta o používateľovi je všeobecne presnejšie porovnávať rovno vlastné dokumenty

Posudok oponenta – otázka 3

- Pre rastúci počet odkrytých dokumentov by mal počet používateľov v tabuľke 10 klesať, ale pre 6 a 8 odkrytých dokumentov sú tieto počty vyššie pri 10 odporúčaných dokumentov, ako je to možné?

Posudok oponenta – odpoveď 3

- Pre 6 až 8 by počet používateľov nemal stúpať a mal byť mať hodnotu 28. Je to preklep ospravedlňujem sa.