

# Algoritmy pro uživatelskou integraci ohodnocené distribuované informace

P. Vojtáš  
Seminař DIS  
Praha, 9.12.2004

DIS 9.12.2004, P. Vojtáš

1

## Vícekritériální vyhledávání

Zákazník hledá levný kvalitní hotel který je blízko určeného místa a ...

Information Retrieval, Web Querying, distribuované, heterogénní, fuzzy

.....  
Pokorný – Vojtáš, ADBIS 2001

Fagin 97, 2001

Fagin, Lotem, Naor 2001

Gunter, Balke, Kiesling 2000

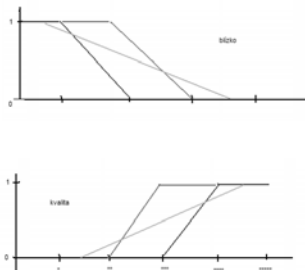
Gurský, Lencses, Vojtáš 2004

.....

DIS 9.12.2004, P. Vojtáš

2

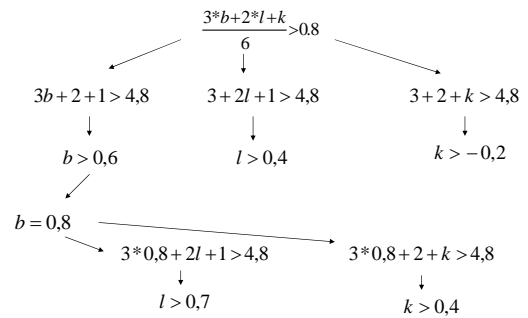
Různí zákazníci – blízko, levný, kvalita, ...



DIS 9.12.2004, P. Vojtáš

3

Počítání s práhem



DIS 9.12.2004, P. Vojtáš

4

Selekce na hodnotu, join s agregovanou hodnotou

blízko			levný			kvalita		
H1	...	0,9	H3	...	0,9	H2	...	0,9
H2	...	0,8	H2	...	0,8	H3	...	0,8
H3	...	0,5	H4	...	0,5	H1	...	0,5
H4	...	0,4	H1	...	0,3	H4	...	0,3

H2 vyhovuje s

$$\frac{3*0,8+2*0,8+0,9}{6}=0,81...$$

DIS 9.12.2004, P. Vojtáš

5

Selekce na hodnotu, join s agregovanou hodnotou

Faginův algoritmus

blízko		levný		kvalita		Zásob.	
H1	0,9	H3	0,9	H2	0,9	H2	0,81
H2	0,8	H2	0,8	H3	0,8	H3	0,68
H3	0,5	H4	0,5	H1	0,5	H1	0,63
H4	0,4	H1	0,3	H4	0,3		

Práh 1

$$(3*0,9+2*0,9+0,9)/6=0,9$$

Ještě neznám nejlepší

Práh 2

$$(3*0,8+2*0,8+0,8)/6=0,8$$

0,81 > 0,8 ... H2 je nejlepší

DIS 9.12.2004, P. Vojtáš

6

Heuristiky – obměny Faginova algoritmu, testy

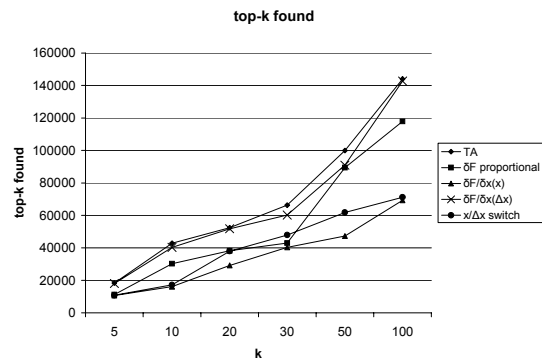
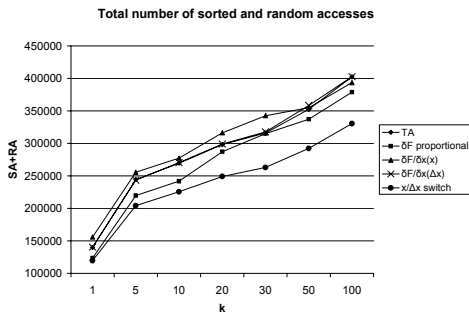
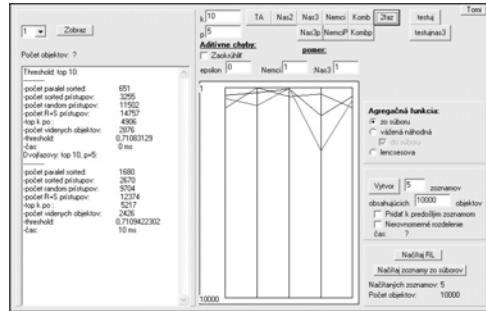
TA = Faginův „threshold algoritmus“ – paralelně, rovnoměrně (vodorovně)

$\Delta F$  - paralelně, proporcionálně k velikosti hodnot F

$\frac{\partial F}{\partial x} * x$  - Největší spád cenové funkce u velikých hodnot (Gurský, Lencses, P.V.)

$\frac{\partial F}{\partial x} * \Delta x$  - Největší spád cenové funkce u rychle klesajících hodnot (Gunter, Balke, Kiesling)

$x / \Delta x$  - sudé kroky – GLV  
- liché kroky – „Němci“



Další experimenty

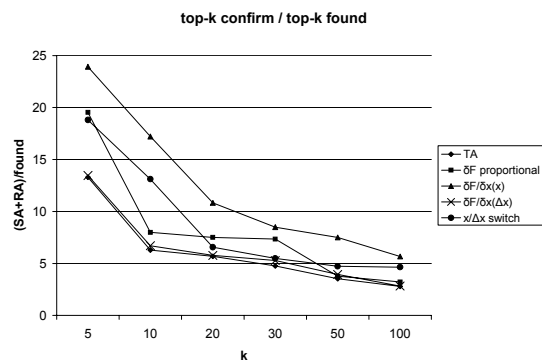
Všechny obměny Faginova „threshold algoritmu“ jsou korektní

Pozorování: nalezení top-k je mnohem rychlejší než potvrzení

Faginův algoritmus je „instance optimal“, až na  $c * m^2$

U rozšíření dotazu – zvětším  $m$  - to už nemusí platit

Distribuce dat  
Učení z předešlých dotazů jiných zákazníků  
Přibližně nejlepší – s aditivní nebo multiplikační chybou



RDF model

**Vize – dotazovací jazyk a implementace pro top-k odpovědí**

Flora nízká\_cena 0,7

DIS 9.12.2004, P. Vojtáš 13

RDF model

[http://apple.cs.umbc.edu/cache/swoogle/N-TRIPLE/59847\\_nt](http://apple.cs.umbc.edu/cache/swoogle/N-TRIPLE/59847_nt)

```
<http://opales.ina.fr/public/eon2003/SimpleReservation>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/2000/01/rdf-schema#Class> .

<http://opales.ina.fr/public/eon2003/meronymyRelation>
<http://www.w3.org/2000/01/rdf-schema#subPropertyOf>
<http://opales.ina.fr/public/eon2003/anyRelation> .

<http://opales.ina.fr/public/eon2003/AnyConcept>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/2000/01/rdf-schema#Class> .

<http://opales.ina.fr/public/eon2003/tripReservationAttribute>
<http://www.w3.org/2000/01/rdf-schema#range>
<http://opales.ina.fr/public/eon2003/AnyConcept> . . . . .
```

DIS 9.12.2004, P. Vojtáš 14

RDF model – from swoogle – většinou „upward“ ontologie, schemata

SimpleReservation	type	Class
meronymyRelation	subPropertyOf	anyRelation
AnyConcept	type	Class
<b>tripReservationAttribute</b>	range	AnyConcept
flightNumber	rdf-schema#label	"flightNumber"@en
internetAvailable	22-rdf-syntax-ns#type	22-rdf-syntax-ns#Property
distanceToBeach	22-rdf-syntax-ns#type	22-rdf-syntax-ns#Property
address	rdf-schema#subPropertyOf	lodgingFacilityAttribute
lodgingChosen	rdf-schema#range	LodgingFacility
flightReservationAttribute	rdf-schema#range	AnyConcept
arrivalCity	rdf-schema#range	City
numberOfBeds	rdf-schema#label	"numberOfBeds"@en

DIS 9.12.2004, P. Vojtáš 15

RDF model pro top-k

Dotazování – SeRQL (typovaný?)

```
Select Hotel, Relevance
From {Hotel} <má_cenu> {nízký?Relevance}
OrderBy Relevance [top-10]
```

Co je potřeba dodat?

Ukládání – horizontální – nevhodné

- vertikální – nevhodné
- class decomposition model – nevhodné
- property decomposition model – nevhodné

Potřebujeme nové ukládání (Sesame nad JDBC a RDBMS)

- indexované podle relevance
- umožňující implementaci např. TA

XML úložiště? Z RDFS do XML Schematu a transformaci...

DIS 9.12.2004, P. Vojtáš 16

Souvislosti s lineární optimalizací

Může LOP pomoci? Konvexní obal dat – pozitivní část – poly algoritmy

Můžou top-k heuristiky pomoci LOP?

Jak je složité spočítat průměty vrcholů simplexů do atributů?

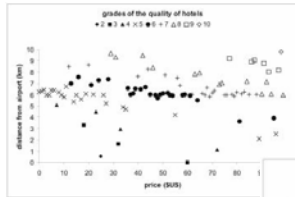
DIS 9.12.2004, P. Vojtáš 17

Různé typy integrace-agregace vícenásobných uživatelských kritérií

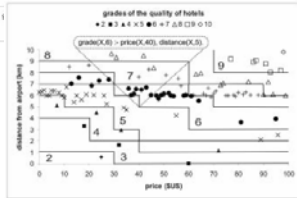
AND – like      OR – like      naučené z dat

DIS 9.12.2004, P. Vojtáš 18

Monotónní klasifikace



Učení z dat



DIS 9.12.2004, P. Vojtáš

19

Děkuji za pozornost, prosím otázky. ...

DIS 9.12.2004, P. Vojtáš

20