Algoritmy pro uživatelskou integraci ohodnocené distribuované informace

P. Vojtáš
Seminař DIS
Praha, 9.12.2004

Vícekriteriální vyhledávání

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

Information Retrieval, Web Querying, distribuované, heterogénní, fuzzy

…..
Pokorný – Vojtáš , ADBIS 2001

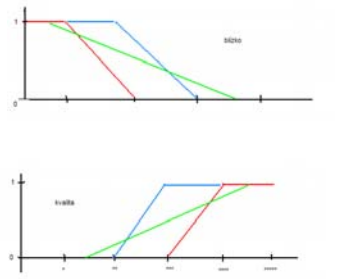
Fagin 97, 2001

Fagin, Lotem, Naor 2001

Guenter, Balke, Kiesling 2000

Gurský, Lencses, Vojtáš 2004
…..

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



Počítání s prahem

$$\frac{3*b+2*l+k}{6} > 0.8$$

$3b+2+1>4,8 \rightarrow b>0,6$

$3+2l+1>4,8 \rightarrow l>0,4$

$3+2+k>4,8 \rightarrow k>-0,2$

$b=0,8$

$3*0,8+2l+1>4,8 \rightarrow l>0,7$

$3*0,8+2+k>4,8 \rightarrow k>0,4$

Selekce na hodnotu, join s agregovanou hodnotou

blízko		
H1	…	0,9
H2	…	0,8
H3	…	0,5
H4	…	0,4

levný		
H3	…	0,9
H2	…	0,8
H4	…	0,5
H1	…	0,3

kvalita		
H2	…	0,9
H3	…	0,8
H1	…	0,5
H4	…	0,3

H2 vyhovuje s $\frac{3*0,8+2*0,8+0,9}{6}=0,81...$

Selekce na hodnotu, join s agregovanou hodnotou

Faginův algoritmus

blízko	
H1	0,9
H2	0,8
H3	0,5
H4	0,4

levný	
H3	0,9
H2	0,8
H4	0,5
H1	0,3

kvalita	
H2	0,9
H3	0,8
H1	0,5
H4	0,3

Zásob.	
H2	0,81
H3	0,68
H1	0,63

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

Heuristiky – obměny Faginova algoritmu, testy

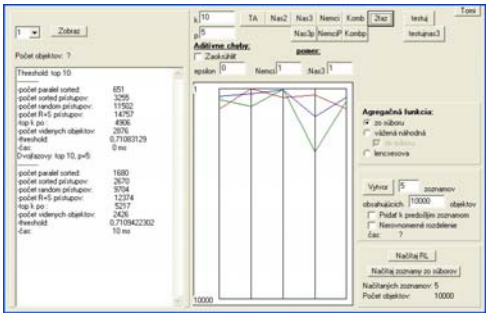
TA = Faginův „threshold algoritmus“ – paralelne, rovnoměrně (vodorovně)

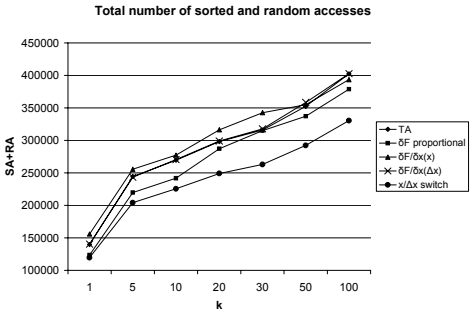
ΔF - paralelně, proporcionálně k velikosti hodnot F

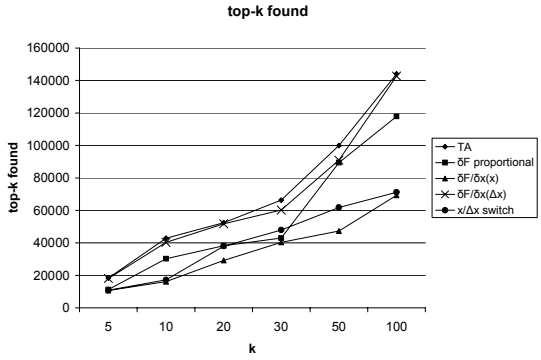
$\frac{\partial F}{\partial x} * x$ - Největší spád cenové funkce u velikých hodnot (Gurský, Lencses, P.V.)

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

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







Další experimenty

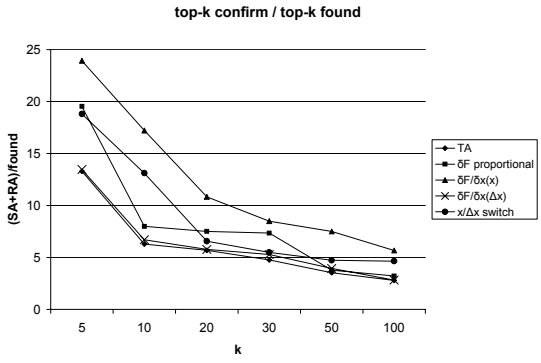
Všechny obměny Faginova „threshold algoritmu “ jsou korektní

Pozorování: nalezení top-k je mnohem rychlejší než potvrzení

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

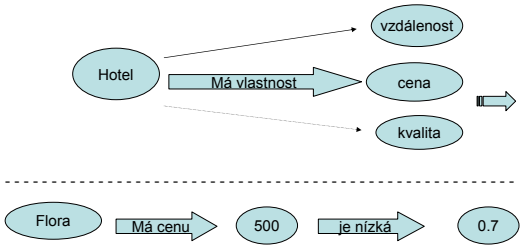
U rozšíření dotazu – zvětším *m* - to už nemusí platit

Distribuce dat
Učení z předešlých dotazů jiných zákazníků
Přibližně nejlepší – s aditivní mebo multiplikativní chybou



RDF model

Vize – dotazovací jazyk a implementace pro top-k odpovědí



Flora nízká_cena 0,7

RDF model

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> . ……

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

SimpleReservation	type	Class
meronymyRelation	subPropertyOf	anyRelation
AnyConcept	type	Class
tripReservationAttribute	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

RDF model pro top-k

Dotazování – SeRQL (typovaný?)

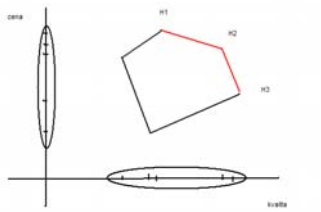
Select Hotel, Relevance
From {Hotel} <má_cenu> {nízký?Relevance}
OrderBy Relevance [top-10]

Co je potřeba dodat?
Ukládání – horizontal – 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...

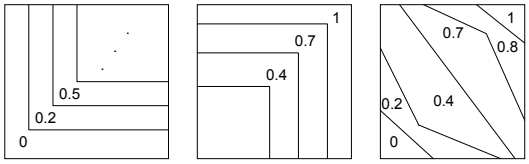
Souvislosti s lineární optimalizací



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

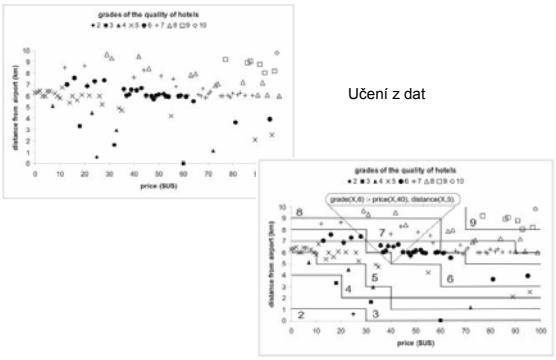
Můžou top-k heuristiky pomoci LOP ?
Jak je složité spočítat průměty vrcholů simplexů do atributů ?

Různé typy integrace-agregace vícenásobných uživatelských kritérií



AND – like OR – like naučené z dat

Monotónní klasifikace



Učení z dat

Děkuji za pozornost, prosím otázky, …