

Thematic Ranking of Object Summaries for Keyword Search

DKE 2018



GEORGIOS J. FAKAS
Department of Information Technology,
Uppsala University,
Sweden

YILUN CAI
Levnovo Group Limited
Hong Kong

ZHI CAI
College of Computer Science
Beijing University of Technology
Beijing, China

NIKOS MAMOULIS
Department of Computer Science
The University of Hong Kong
Hong Kong

Outline

1. Motivation

2. Background & Related work

3. Themtiac Size- l OSs

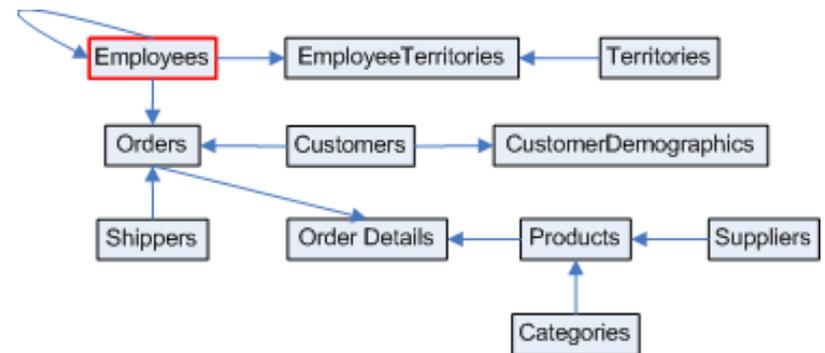
4. Approaches

5. Evaluation Results

6. Conclusion & Future Work

1.1 Object Summaries

- Relational Databases are everywhere: Web, Desktops etc.
- Social graphs are also everywhere!
- Difficult to retrieve information about a **Data Subject** (DS) unless you know very well:
 - **SQL** and
 - **Schema details** etc.
- There is a need for keyword search facilities analogous to Web

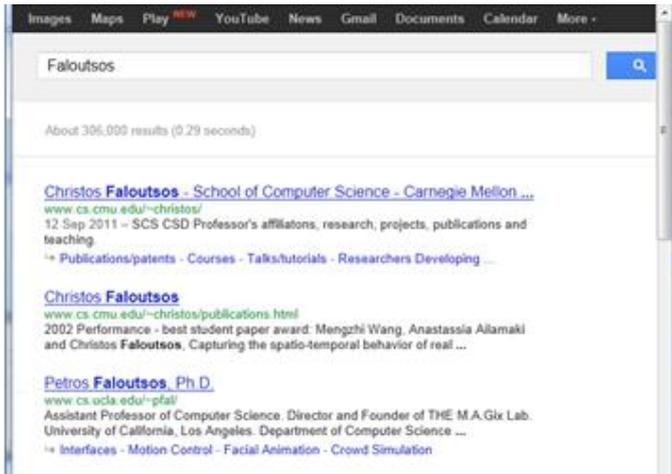


```
Select *  
From Employees, Orders, Shippers  
Where Employees.ID=Orders.ID  
AND Orders.Shipper=Shippers.ID  
AND Name="Leverling"
```

1.1 Object Summaries

Query Search: **Faloutsos**

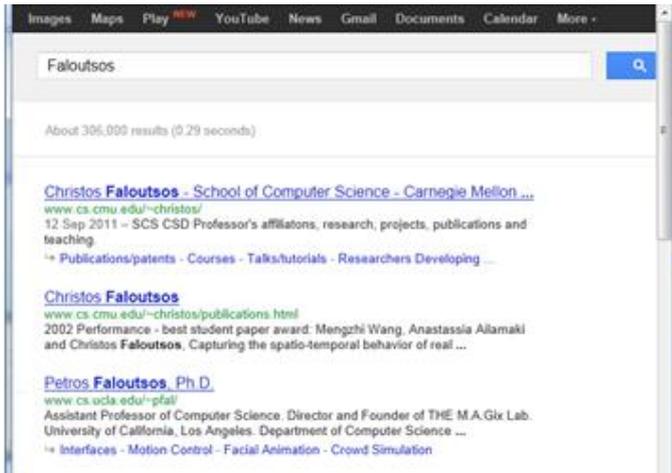
Web Search Result: **ranked set of links**
and snippets



1.1 Object Summaries

Query Search: **Faloutsos**

Web Search Result: **ranked set of links**
and snippets



Images Maps Play ^{NEW} YouTube News Gmail Documents Calendar More

Faloutsos

About 306,000 results (0.29 seconds)

[Christos Faloutsos - School of Computer Science - Carnegie Mellon ...](#)
www.cs.cmu.edu/~christos/
12 Sep 2011 - SCS CSD Professor's affiliations, research, projects, publications and teaching.
+ Publications/patents - Courses - Talks/tutorials - Researchers Developing ...

[Christos Faloutsos](#)
www.cs.cmu.edu/~christos/publications.html
2002 Performance - best student paper award. Mengzhi Wang, Anastassia Ailamaki and Christos Faloutsos, Capturing the spatio-temporal behavior of real ...

[Petros Faloutsos, Ph.D.](#)
www.cs.ucla.edu/~pfal/
Assistant Professor of Computer Science. Director and Founder of THE M.A.GiX Lab. University of California, Los Angeles. Department of Computer Science ...
+ Interfaces - Motion Control - Facial Animation - Crowd Simulation

[\[Publications/patents\]](#) [\[Software\]](#) [\[Talks/tutorials\]](#) [\[Courses\]](#) [\[Service\]](#) [\[Misc.\]](#)



U.S. mail address:
Christos Faloutsos
[Dept. of Computer Science](#), GHC 8019
[Carnegie Mellon University](#)
5000 Forbes Avenue
Pittsburgh, PA 15213-3891
Admin: Marilyn Walgora
412-268.3505, GHC 8120
email [first-initial-and-her-last-name] AT cs DOT cmu DOT edu

Christos Faloutsos

Current Position: Professor.
Courtesy appointment: [Electrical and Computer Engineering, CMU](#)

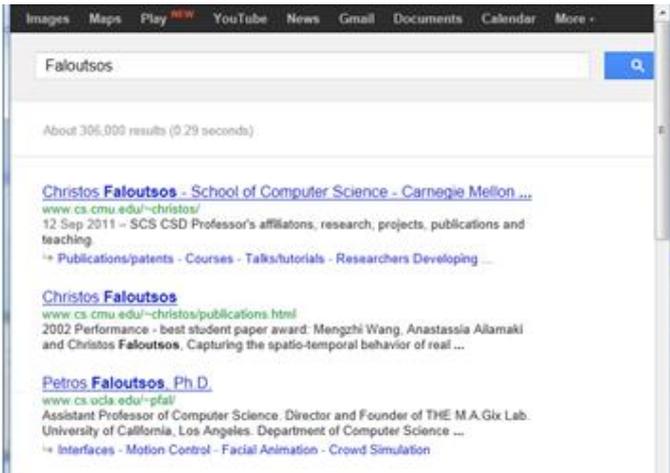
1.1 Object Summaries

Query Search: **Faloutsos**

Web Search Result: **ranked set of links and snippets**

Query Search: **Faloutsos**

OS Result: **set of OSs and size-1 OSs.**



1.1 Object Summaries

Query Search: **Faloutsos**

Web Search Result: **ranked set of links and snippets**

A screenshot of a web search engine interface. The search bar contains the text 'Faloutsos'. Below the search bar, it indicates 'About 306,000 results (0.29 seconds)'. The search results are listed as follows:

- Christos Faloutsos - School of Computer Science - Carnegie Mellon ...**
www.cs.cmu.edu/~christos/
12 Sep 2011 - SCS CSD Professor's affiliations, research, projects, publications and teaching.
+ Publications/patents - Courses - Talks/tutorials - Researchers Developing ...
- Christos Faloutsos**
www.cs.cmu.edu/~christos/publications.html
2002 Performance - best student paper award: Mengzhi Wang, Anastassia Ailamaki and Christos Faloutsos, Capturing the spatio-temporal behavior of real ...
- Petros Faloutsos, Ph.D.**
www.cs.ucla.edu/~pfall/
Assistant Professor of Computer Science. Director and Founder of THE M.A.GiX Lab. University of California, Los Angeles. Department of Computer Science ...
+ Interfaces - Motion Control - Facial Animation - Crowd Simulation

A screenshot of a personal profile page for Christos Faloutsos. The page includes a navigation menu with links: [Publications/patents] [Software] [Talks/tutorials] [Courses] [Service] [Misc]. Below the menu is a portrait photo of Christos Faloutsos. To the right of the photo, the following information is provided:

U.S. mail address:
Christos Faloutsos
Dept. of Computer Science, GHC 8019
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213-3891
Admin: Marilyn Walgora
412-268.3505, GHC 8120
email [first-initial-and-her-last-name] AT cs DOT cmu DOT edu

Christos Faloutsos

Current Position: Professor.
Courtesy appointment: [Electrical and Computer Engineering, CMU](#)

Query Search: **Faloutsos**

OS Result: **set of OSs and size-1 OSs.**

A screenshot of the Object Summaries (OS) interface. The title 'Object Summaries (OS)' is displayed in a large, colorful font. Below the title, the subtitle reads 'A Novel Keyword Search Paradigm in Relational Databases'. A search bar contains the text 'Faloutsos' and a 'Search' button. Below the search bar, there are links for 'Advanced Search', 'Affinity Calculator', and 'Global Importance Score New!'. At the bottom of the search bar area, there are links: 'About OSs - About Demo - People - Publications'.

Object Summaries (OS) Christos Faloutsos Search Advanced Search Affinity Calculator

DBLP: search for Christos Faloutsos

Author: Christos Faloutsos

Paper: The QBIC Project: Querying Images by Content, Using Color, Texture, and Shape.
Co-Author: Dragutin Petrovic Gabriel Tauben Wayne Niblack Myron Flickner Ron Barber Peter Yanker William Equitz Eduardo H. Glasman

Conference: Storage and Retrieval for Image and Video Databases (SPIE) 1993
Cited by: Organization and Retrieval in a Pictorial Digital Library.
Author: Yuri Quntana
Conference: ACM DL 1997

Cited by: Merging Ranks from Heterogeneous Internet Sources.
Author: Hector Garcia-Molina Luis Gravano
Conference: VLDB 1997

Cited by: Symbolic Description and Visual Querying of Image Sequences Using Spatio-Temporal Logic.
Author: Alberto Del Bimbo Enrico Vicario Daniele Zingoni
Conference: IEEE Trans. Knowl. Data Eng. 1995

Cited by: Optimizing Queries Across Diverse Data Sources.
Author: Jun Yang Donald Kossmann Laura M. Haas Edward L. Wimmers
Conference: VLDB 1997

Outline

1. Motivation

2. Background & Related work

3. Themtiac Size-*l* OSs

4. Approaches

5. Evaluation Results

6. Conclusion & Future Work

2.1 Object Summaries

OS Generation - Methodology

KW-ID = "Janet Leverling"

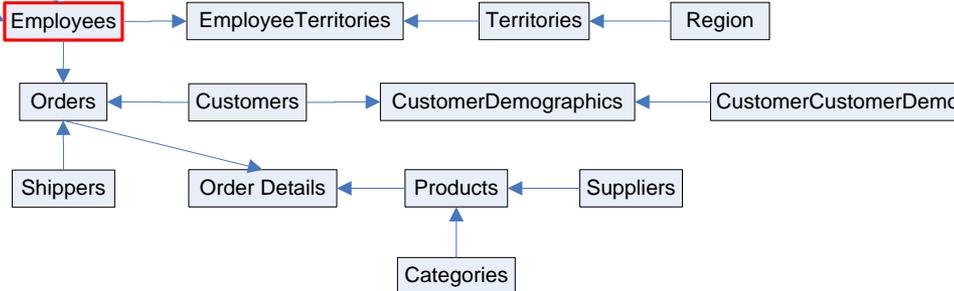
Territories				Region	
TerritoryID	TerritoryDescription	RegionID	RegionID	RegionDescription	
t1	Rockville	1	r1	Eastern	
t2	Greensboro	1	r4	Southern	
t3	Cary	1			
t4	Atlanta	4			

Employees							EmployeeTerritories	
EmployeeID	LastName	FirstName	Title	TitleOfCourtesy	Address	ReportsTo	EmployeeID	TerritoryID
e1	Davolio	Nancy	Sales Representative	Ms.	507 - 20th Ave. E. Apt. 2A 2		et1	30346
e2	Fuller	Andrew	Vice President, Sales Dr.		908 W. Capital Way	NULL	et2	20852
e3	Leverling	Janet	Sales Representative	Mrs.	722 Moss Bay Blvd.	2	et3	27403
e4	Peacock	Margaret	Sales Representative	Mrs.	4110 Old Redmond Rd.	2	et4	27511

Orders							Customers				Shippers			
OrderID	CustomerID	EmployeeID	ShipVia	ShipName	ShipAddress		CustomerID	CompanyName	ContactName	Address	ShipperID	CompanyName	Phone	
o1	10258	ERNSH	1	1	Ernst Handel	Kirchgasse 6	c1	ERNSH	Ernst Handel	Roland Mendel	Kirchgasse 6	s1	Speedy Express	(503) 555-9831
o2	10273	QUICK	3	3	QUICK-Stop	Taucherstraße 10	c2	QUICK	QUICK-Stop	Margaret Peacock	Taucherstraße 10	s2	United Package	(503) 555-3199
o3	10285	QUICK	1	2	QUICK-Stop	Taucherstraße 10	c3	SAVEA	Save-a-lot Markets	Horst Kloss	187 Suffolk Ln.	s3	Federal Shipping	(503) 555-9931
o4	10393	SAVEA	1	3	Save-a-lot Markets	187 Suffolk Ln.								
o5	10398	QUICK	2	3	QUICK-Stop	Taucherstraße 10								
o6	10418	QUICK	4	1	QUICK-Stop	Taucherstraße 10								
o7	10451	SAVEA	4	3	Save-a-lot Markets	187 Suffolk Ln.								

Order Details					Products					Suppliers					Categories				
OrderID	ProductID	UnitPrice	Quantity	Discount	ProductID	ProductName	SupplierID	CategoryID	QuantityPerUnit	UnitsInStock	SupplierID	CompanyName	ContactName	ContactTitle	Address	City	CategoryID	CategoryName	Description
od1	10273	2	15.2000	50	0.2	p1	Chai	1	1	10 boxes x 20 bags	39	su1	Exotic Liquids	Charlotte Cooper	Purchasing Manager	49 Gilbert St. London	ca1	Beverages	Soft drinks, coffees, teas, beers, ...
od2	10285	2	15.2000	25	0.25	p2	Chang	1	1	24 - 12 oz bottles	17								
od3	10398	2	14.4000	30	0.0														
od4	10418	1	7.6000	55	0.0														
od5	10418	2	15.2000	60	0.0														
od6	10451	2	19.2000	120	0.1														

- t^{DS} a central tuple containing the Kw; tuples around t^{DS} contain additional information about the Data Subject.
- R^{DS} the corresponding central Relation; similarly Relations around contain additional information.

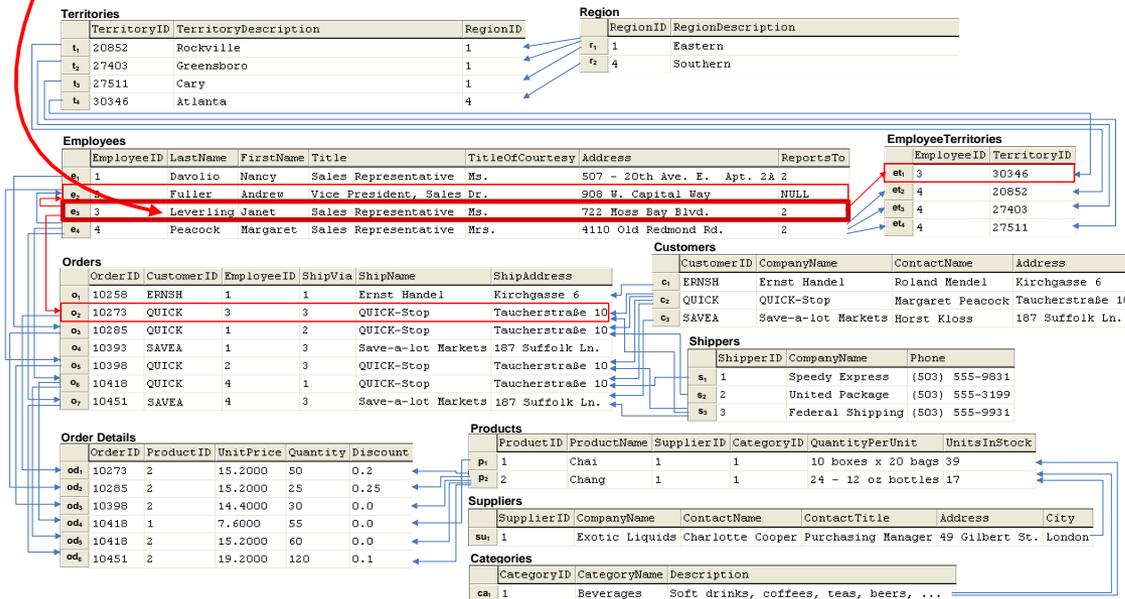


[Fakas, DKE, 2011]

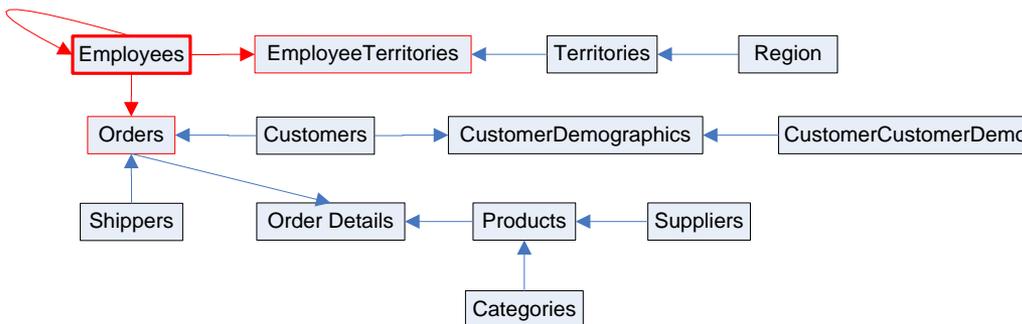
2.1 Object Summaries

OS Generation - Methodology

KW-ID = "Janet Leverling"



- t^{DS} a central tuple containing the Kw; tuples around t^{DS} contain additional information about the Data Subject.
- R^{DS} the corresponding central Relation; similarly Relations around contain additional information.



[Fakas, DKE, 2011]

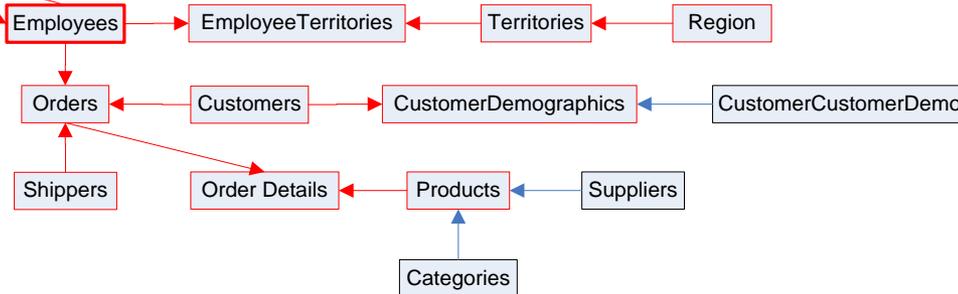
2.1 Object Summaries

OS Generation - Methodology

KW-ID = "Janet Leverling"

The diagram illustrates the relationships between various database tables. Red boxes highlight specific data points, and blue arrows show the flow of relationships between them. The tables include Territories, Region, Employees, EmployeeTerritories, Orders, Customers, Shippers, Order Details, Products, Suppliers, and Categories.

- t^{DS} a central tuple containing the Kw; tuples around t^{DS} contain additional information about the Data Subject.
- R^{DS} the corresponding central Relation; similarly Relations around contain additional information.



[Fakas, DKE, 2011]

2.1 Object Summaries

OS Generation - Methodology

KW-ID = "Janet Leverling"

Territories				Region	
TerritoryID	TerritoryDescription	RegionID	RegionID	RegionDescription	
t ₁	20852 Rockville	1	r ₁	1 Eastern	
t ₂	27403 Greensboro	1	r ₂	4 Southern	
t ₃	27511 Cary	1			
t ₄	30346 Atlanta	4			

Employees						EmployeeTerritories		
EmployeeID	LastName	FirstName	Title	TitleOfCourtesy	Address	ReportsTo	EmployeeID	TerritoryID
e ₁	Davolio	Nancy	Sales Representative	Ms.	507 - 20th Ave. E. Apt. 2A 2		et ₁	3 30346
e ₂	Fuller	Andrew	Vice President, Sales Dr.		908 U. Capital Way	NULL	et ₂	4 20852
e ₃	Leverling	Janet	Sales Representative	Ms.	722 Moss Bay Blvd.	2	et ₃	4 27403
e ₄	Peacock	Margaret	Sales Representative	Mrs.	4110 Old Redmond Rd.		et ₄	4 27511

Orders					
OrderID	CustomerID	EmployeeID	ShipVia	ShipName	ShipAddress
o ₁	10258	ERNSH	1	Ernst Handel	Kirchgasse 6
o ₂	10273	QUICK	3	QUICK-Stop	Taucherstraße 10
o ₃	10285	QUICK	1	QUICK-Stop	Taucherstraße 10
o ₄	10393	SAVEA	1	Save-a-lot Markets	187 Suffolk Ln.
o ₅	10398	QUICK	2	QUICK-Stop	Taucherstraße 10
o ₆	10418	QUICK	4	QUICK-Stop	Taucherstraße 10
o ₇	10451	SAVEA	4	Save-a-lot Markets	187 Suffolk Ln.

Customers			
CustomerID	CompanyName	ContactName	Address
c ₁	ERNSH	Ernst Handel	Roland Mendel Kirchgasse 6
c ₂	QUICK	QUICK-Stop	Margaret Peacock Taucherstraße 10
c ₃	SAVEA	Save-a-lot Markets	Horst Kloss 187 Suffolk Ln.

Shippers		
ShipperID	CompanyName	Phone
s ₁	Speedy Express	(503) 555-9831
s ₂	United Package	(503) 555-3199
s ₃	Federal Shipping	(503) 555-9931

Products					
ProductID	ProductName	SupplierID	CategoryID	QuantityPerUnit	UnitsInStock
p ₁	Chai	1	1	10 boxes x 20 bags 39	
p ₂	Chang	1	1	24 - 12 oz bottles 17	

Suppliers					
SupplierID	CompanyName	ContactName	ContactTitle	Address	City
su ₁	Exotic Liquids	Charlotte Cooper	Purchasing Manager	49 Gilbert St.	London

Categories		
CategoryID	CategoryName	Description
ca ₁	Beverages	Soft drinks, coffees, teas, beers, ...

OS for "Janet Leverling"

Employees					
EmployeeID	LastName	FirstName	Title	Address	PostalCode
3	Leverling	Janet	Sales Representative	722 Moss Bay Blvd.	98033 ... (e ₃)

Employees (Reports To)	
LastName	FirstName
Fuller	Andrew (e ₂)

Territories, Region	
TerritoryDescription	RegionDescription
Atlanta	Southern (et ₁ , t ₂)

Orders				
OrderID	ShipName	ShipAddress	OrderDate	RequiredDate
10273	QUICK-Stop	Taucherstraße 10	1996-08-05	1996-09-02 00 1996-08-12 (o ₂)

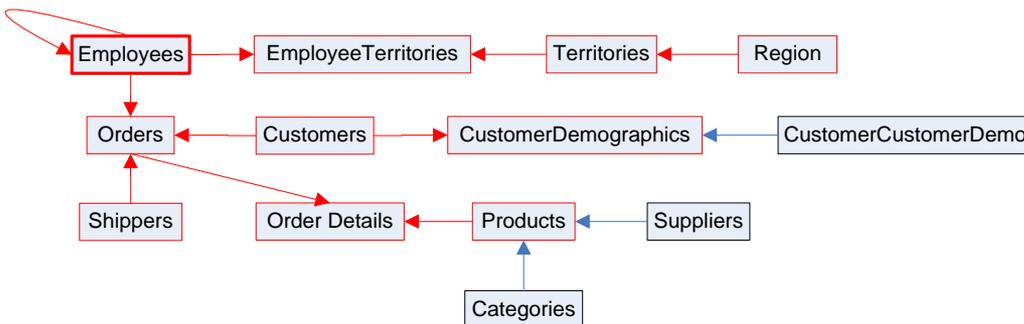
Customers	
CompanyName	ContactName
QUICK-Stop	Margaret Peacock (c ₂)

Shippers
CompanyName
Federal Shipping (s ₃)

Order Details			
UnitPrice	Quantity	Discount	
15.2000	50	0.2 (od ₁)	

Products	
ProductName	QuantityPerUnit
Chang	24 - 12 oz bottles (p ₂)

Categories	
CategoryName	Description
Beverages	Soft drinks, coffees, teas, beers, ... (ca ₁)



2.3 Motivation

Ranking of Size-l OSs

Query: *identifying keyword: Chen*

For the DBLP dataset, there 1,982 OSs, i.e. 1,982 authors having the name “Chen”.

Using Authoritative ranking, Peter Chen will always be ranked first because of his many citations. This is ineffective for users who search for a DS that does not have the best importance scores.

In view of this, in this paper, we propose the thematic ranking of OSs, where thematic keywords are also input by the user.

2.3 Motivation

Thematic Ranking of Size-l OSs

Query: *identifying keyword*: **Chen**

thematic keyword: **Mining**

the additional thematic keyword makes '**Ming-Syan Chen**' prevail, since his OS contains '**Mining**' many times.

Author: Ming-Syan Chen [1.00, 0.45]
Paper: A robust and efficient clustering algorithm based...
Co-Author: Cheng-Ru Lin. Conf.: KDD. Year: 2002
Paper: Distributed data *mining* in a chain store... [0.98, 0.16]
Co-Author: Philip S. Yu, ... Conf.: KDD. Year: 2002.
Paper: *Mining* Relationship between Triggering... [0.98, 0.15]
Co-Author: Philip S. Yu, ... Conf.: SDM. Year: 2002.
Paper: DOMISA: DOM-Based infor... *Mining*... [0.98, 0.18]
Co-Author: Hung-Yu Kao, ... Conf.: SDM, Year: 2004.
Paper: Efficient ... *Mining* for Association Rules.. [0.98, 0.19]
Co-Author: Philip S. Yu, ... Conf.: CIKM Year: 1995.
Cited by: Parallel *Mining* of Association... [0.93, 0.36]
Cited by: Data *Mining*: An Overview from... [0.93, 0.82]
Cited by: Dynamic Load Balan.. *Mining*... [0.93, 0.16]
Cited by: Parallel *Mining* Algorithms for G... [0.93, 0.16]

.....

Outline

- 1. Motivation**
- 2. Background & Related work**
- 3. Themtiac Size-*l* OSs**
- 4. Approaches**
- 5. Evaluation Results**
- 6. Conclusion & Future Work**

3 Thematic Size-1 OSs

Author: Ming-Syan Chen [1.00, 0.45]
Paper: A robust and efficient clustering algorithm based...
Co-Author: Cheng-Ru Lin. **Conf.:** KDD. **Year:** 2002
Paper: Distributed data mining in a chain store... [0.98, 0.16]
Co-Author: Philip S. Yu, ... **Conf.:** KDD. **Year:** 2002.
Paper: Mining Relationship between Triggering...[0.98, 0.15]
Co-Author: Philip S. Yu, ... **Conf.:** SDM. **Year:** 2002.
Paper: DOMISA: DOM-Based infor... Mining... [0.98, 0.18]
Co-Author: Hung-Yu Kao, ... **Conf.:** SDM, **Year:** 2004.
Paper: Efficient ... Mining for Association Rules..[0.98, 0.19]
Co-Author: Philip S. Yu, ... **Conf.:** CIKM **Year:** 1995.
Cited by: Parallel Mining of Association... [0.93, 0.36]
Cited by: Data Mining: An Overview from... [0.93, 0.82]
Cited by: Dynamic Load Balan.. Mining... [0.93, 0.16]
Cited by: Parallel Mining Algorithms for G... [0.93, 0.16]

.....

Definition:

A query Q comprises **two sets of keywords**, $Q = \langle q_1, q_2 \rangle$,

- q_1 is a set of **identifying keywords**
- q_2 is a set of **thematic keywords**

Criteria:

1. global Importance;
2. IR-properties; and
3. Affinity

3 Thematic Size-1 OSs

Author: Ming-Syan Chen [1.00, 0.45]
Paper: A robust and efficient clustering algorithm based...
Co-Author: Cheng-Ru Lin. **Conf.:** KDD. **Year:** 2002
Paper: Distributed data mining in a chain store... [0.98, 0.16]
Co-Author: Philip S. Yu, ... **Conf.:** KDD. **Year:** 2002.
Paper: Mining Relationship between Triggering...[0.98, 0.15]
Co-Author: Philip S. Yu, ... **Conf.:** SDM. **Year:** 2002.
Paper: DOMISA: DOM-Based infor... Mining... [0.98, 0.18]
Co-Author: Hung-Yu Kao, ... **Conf.:** SDM, **Year:** 2004.
Paper: Efficient ... Mining for Association Rules..[0.98, 0.19]
Co-Author: Philip S. Yu, ... **Conf.:** CIKM **Year:** 1995.
Cited by: Parallel Mining of Association... [0.93, 0.36]
Cited by: Data Mining: An Overview from... [0.93, 0.82]
Cited by: Dynamic Load Balan.. Mining... [0.93, 0.16]
Cited by: Parallel Mining Algorithms for G... [0.93, 0.16]

Definition:

A query Q comprises **two sets of keywords**, $Q = \langle q_1, q_2 \rangle$,

- q_1 is a set of **identifying keywords**
- q_2 is a set of **thematic keywords**

Criteria:

1. global Importance;
2. IR-properties; and
3. Affinity

$$score_1(O, q_1) = Im(t^{DS}),$$

where

$$score_2(O, q_2) = \frac{\sum_{t \in O} s(t, q_2)}{1 - \alpha + \alpha \cdot \frac{dl(O)}{avdl(OS)}}$$

$$s(t, q_2) = \sum_{w \in t \cap q_2} (1 + \ln(1 + \ln(tf_w(t)))) \cdot \ln(idf_w) \cdot li(t),$$

$$idf_w = \frac{N_{OS} + 1}{df_w(OS)},$$

$$score(O, Q) = score_1(O, q_1) \cdot score_2(O, q_2)$$

$$li(t) = Af(t) \cdot Im(t).$$

Outline

1. Motivation

2. Related work

3. Themtiac Size-*l* OSs

4. Approaches

5. Evaluation Results

6. Conclusion & Future Work

4.1 Problem reformulation

We reformulate our OSs ranking problem as a *top-k* Group By join problem (**kGBJ**).

Considering two selection operations on R^{DS} and R^{TH} then we get $R^{DS}(q_1)$ and $R^{TH}(q_2)$

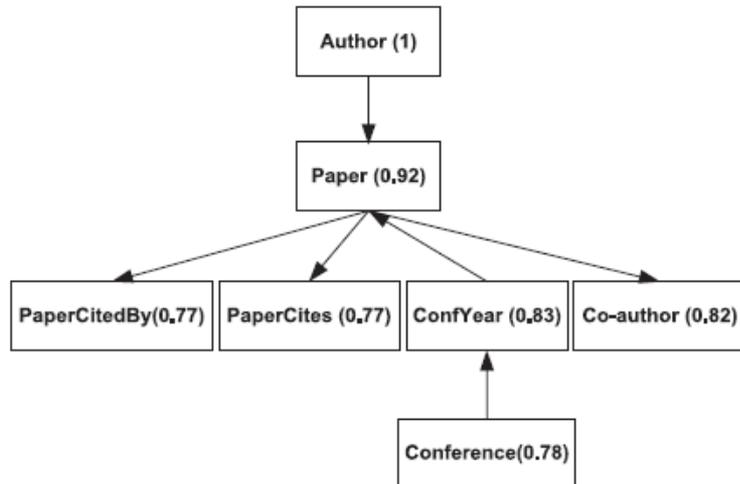


Fig. 2. DBLP Author G^{DS} (Affinity).

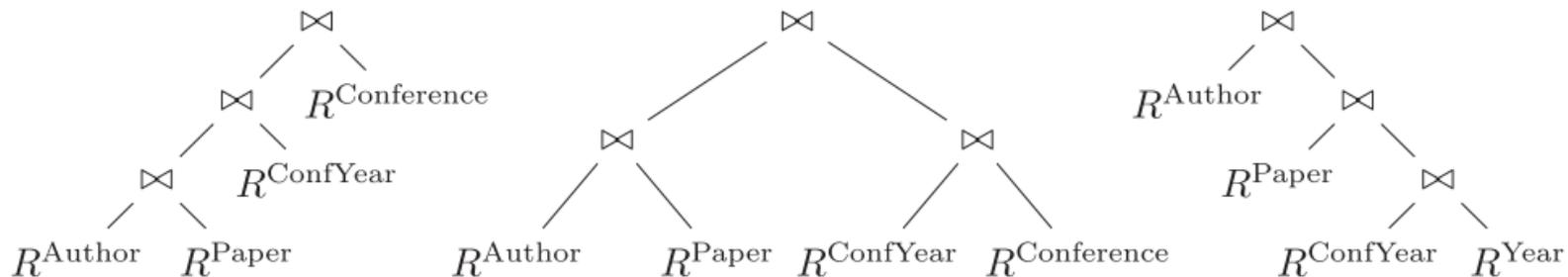
Identifying Keywords (q_1)	Frequency in DBLP
David	4,235
Chen	1,982
Wang	1,778
Alan	660
John	3,717
Nick	179

Thematic Keywords (q_2)	Frequency in DBLP
Mining	2,961
System	32,253
Logic	65
Data	22,500

4.1 Baseline: Bi-Directional approach

BD approach

As in a query optimizer, given the sizes of $R^{DS}(q_1)$ and $R^{TH}(q_2)$, the estimation of the optimal **meeting point** is done with the help of statistics.



Meeting point Examples

4.2 Top- k Bi-Directional approach

kBD approach

Rationale of this approach is to avoid the entire BD traversal and processing of our input (i.e. of $R^{DS}(q_1)$ and $R^{TH}(q_2)$).

We achieve this by estimating **upper** and **lower bounds** for each OS and by managing them in descending order of their upper bounds in a **max-heap**.

Outline

- 1. Motivation**
- 2. Background & Related work**
- 3. Themtiac Size-*l* OSs**
- 4. Approaches**
- 5. Evaluation Results**
- 6. Conclusion & Future Work**

5. Experimental Evaluation

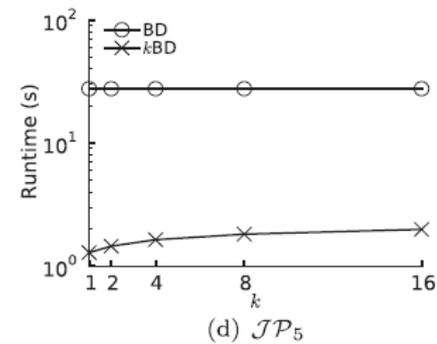
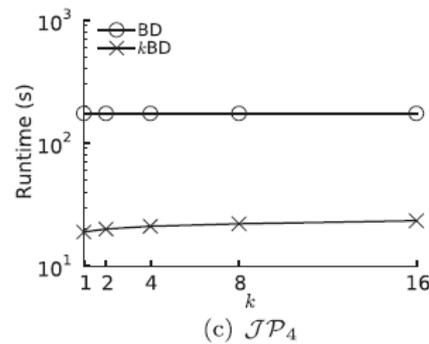
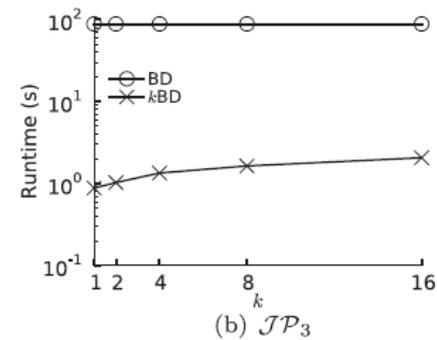
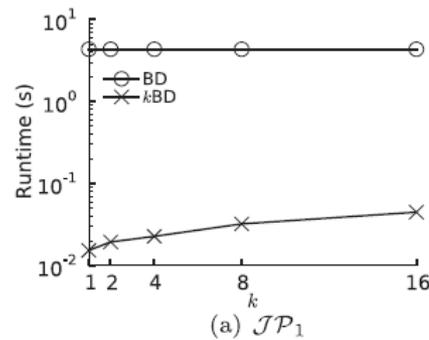
Effectiveness

k	Precision(=Recall)	Ranking Correlation
5	92.0%	0.84
10	96.5%	0.92
15	98.8%	0.96
20	100%	0.98
25	100%	0.99

Precision(Recall) and Ranking Correlation

5. Experimental Evaluation

Efficiency



**Efficiency of BD and kBD
for Various Values of k**

Outline

- 1. Motivation**
- 2. Background & Related work**
- 3. Themtiac Size-*l* OSs**
- 4. Approaches**
- 5. Evaluation Results**
- 6. Conclusion & Future Work**

6.1 Conclusion & Future Work

Contributions

- The formal definition of *thematic ranking object summaries* for keyword search.
- The an efficient *top-k group-by join* algorithm.
- **Applications:** Google, Google Desktop, DBMS, etc.

Thematic Ranking of Object Summaries for Keyword Search



Thank you

Questions!

4.2 Top- k Bi-Directional approach

kBD approach

Upper and **lower** bounds of OSs are calculated as follows:

$$LB(O) = In(O) \cdot \sum_{j=1}^l n_j \cdot s(t_j)$$

$$UB_1(O, Q) = LB(O) + In(O) \cdot (M - \sigma) \cdot s(t_{l+1})$$

$$\gamma = \min\{m \cdot (|R^{Th}(q_2)| - l), M - \sigma\}$$

tighter upper bound

$$UB_2(O, Q) = LB(O) + In(O) \cdot \left(\sum_{j=1}^d s(t_{l+j}) \cdot m + s(t_{l+d+1}) \cdot r \right) \quad d = \lfloor \gamma/m \rfloor \quad r = \gamma \bmod m$$

further tighten

$$UB_3(O, Q) = LB(O) + In(O) \cdot \left(\sum_{j=1}^d m \cdot s(t_{a_j}) + r \cdot s(t_{a_{d+1}}) \right)$$

4.2 Top- k Bi-Directional approach

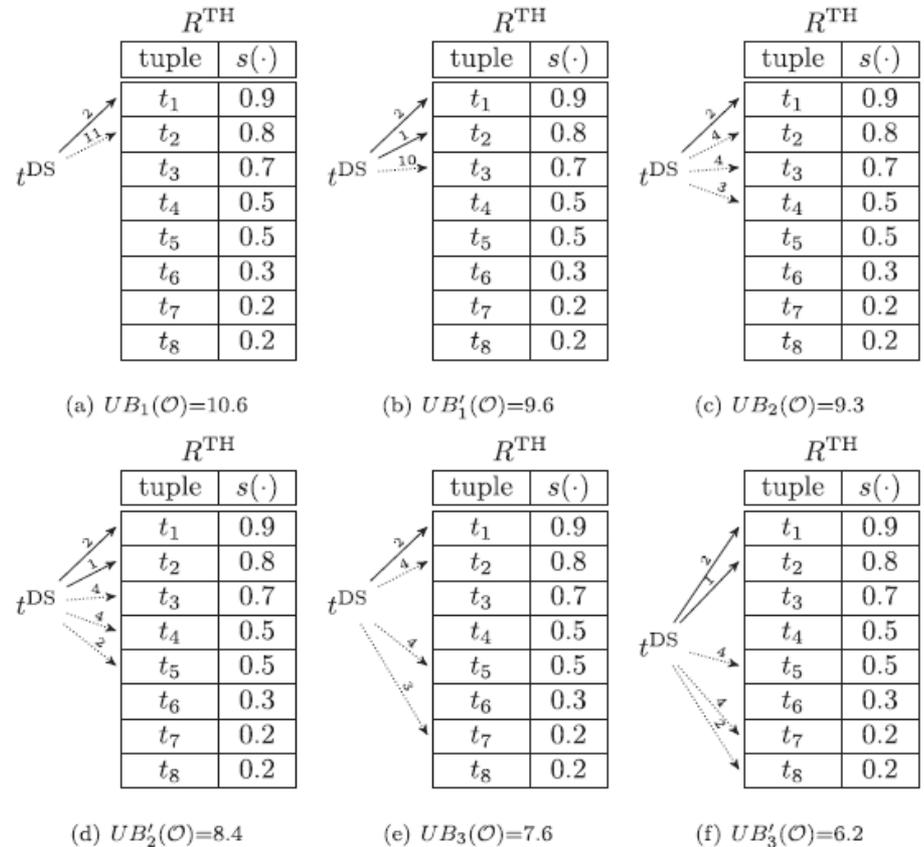
kBD approach

Algorithm 1. kBD Algorithm

$kBD(R^{DS}(q_1), R^{Th}(q_2), k)$

- 1: $H := \emptyset$;
- 2: $L^{Th} := R^{Th}(q_2)$;
- 3: sort tuples in L^{Th} in descending order of their $s(\cdot)$ scores;
- 4: **for** each \mathcal{O} w.r.t. t^{DS} in $R^{DS}(q_1)$ **do**
- 5: $LB(\mathcal{O}) := 0$; $UB(\mathcal{O}) := \text{CALCUB } \mathcal{O}$;
- 6: insert \mathcal{O} into H with priority $UB(\mathcal{O})$;
- 7: **while** $k > 0 \wedge H$ is not empty **do**
- 8: pop \mathcal{O}_{cur} from H ;
- 9: $\mathcal{O}_{next} := H.top()$;
- 10: **if** $LB(\mathcal{O}_{cur}) \geq UB(\mathcal{O}_{next})$ **then**
- 11: report \mathcal{O}_{cur} as a result;
- 12: $k := k - 1$;
- 13: **else**
- 14: $t_i :=$ the next tuple in L^{Th} can join with \mathcal{O}_{cur} ;
- 15: $n := \text{JOIN } \mathcal{O}_{cur}, t_i$;
- 16: $LB(\mathcal{O}_{cur}) := LB(\mathcal{O}_{cur}) + n \cdot In(\mathcal{O}_{cur}) \cdot s(t_i, q_2)$;
- 17: $UB(\mathcal{O}_{cur}) := \text{CALCUB } \mathcal{O}_{cur}$;
- 18: push \mathcal{O}_{cur} back into H with priority $UB(\mathcal{O}_{cur})$;

Calculating the Upper Bound Scores of an OS \mathcal{O} ($In(\mathcal{O}) = 1.0$, $M = 13$, $m = 4$)



4.2 Top- k Bi-Directional approach

kBD approach

Algorithm 1. k BD Algorithm

k BD ($R^{DS}(q_1), R^{Th}(q_2), k$)

- 1: $H := \emptyset$;
- 2: $L^{Th} := R^{Th}(q_2)$;
- 3: sort tuples in L^{Th} in descending order of their $s(\cdot)$ scores;
- 4: **for** each \mathcal{O} w.r.t. t^{DS} in $R^{DS}(q_1)$ **do**
- 5: $LB(\mathcal{O}) := 0$; $UB(\mathcal{O}) := \text{CALCUB } \mathcal{O}$;
- 6: insert \mathcal{O} into H with priority $UB(\mathcal{O})$;
- 7: **while** $k > 0 \wedge H$ is not empty **do**
- 8: pop \mathcal{O}_{cur} from H ;
- 9: $\mathcal{O}_{next} := H.\text{top}()$;
- 10: **if** $LB(\mathcal{O}_{cur}) \geq UB(\mathcal{O}_{next})$ **then**
- 11: report \mathcal{O}_{cur} as a result;
- 12: $k := k - 1$;
- 13: **else**
- 14: $t_i :=$ the next tuple in L^{Th} can join with \mathcal{O}_{cur} ;
- 15: $n := \text{JOIN } \mathcal{O}_{cur}, t_i$;
- 16: $LB(\mathcal{O}_{cur}) := LB(\mathcal{O}_{cur}) + n \cdot \text{In}(\mathcal{O}_{cur}) \cdot s(t_i, q_2)$;
- 17: $UB(\mathcal{O}_{cur}) := \text{CALCUB } \mathcal{O}_{cur}$;
- 18: push \mathcal{O}_{cur} back into H with priority $UB(\mathcal{O}_{cur})$;

The k BD Algorithm for $k = 1$

OS	$UB(\cdot)$	$LB(\cdot)$
\mathcal{O}_1	8.0	0
\mathcal{O}_2	6.0	0
\mathcal{O}_3	10.0	0
\mathcal{O}_4	5.0	0
\mathcal{O}_5	4.0	0

$H = \langle \mathcal{O}_3, \mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_4, \mathcal{O}_5 \rangle$,
 $\mathcal{O}_{cur} = \mathcal{O}_3, \mathcal{O}_{next} = \mathcal{O}_1$
 (a) Initialization

OS	$UB(\cdot)$	$LB(\cdot)$
\mathcal{O}_1	8.0	0
\mathcal{O}_2	6.0	0
\mathcal{O}_3	7.0	6.5
\mathcal{O}_4	5.0	0
\mathcal{O}_5	4.0	0

$H = \langle \mathcal{O}_1, \mathcal{O}_3, \mathcal{O}_2, \mathcal{O}_4, \mathcal{O}_5 \rangle$,
 $\mathcal{O}_{cur} = \mathcal{O}_1, \mathcal{O}_{next} = \mathcal{O}_3$
 (b) Iteration 1

OS	$UB(\cdot)$	$LB(\cdot)$
\mathcal{O}_1	7.5	5.0
\mathcal{O}_2	6.0	0
\mathcal{O}_3	7.0	6.5
\mathcal{O}_4	5.0	0
\mathcal{O}_5	4.0	0

$H = \langle \mathcal{O}_1, \mathcal{O}_3, \mathcal{O}_2, \mathcal{O}_4, \mathcal{O}_5 \rangle$,
 $\mathcal{O}_{cur} = \mathcal{O}_1, \mathcal{O}_{next} = \mathcal{O}_3$
 (c) Iteration 2

OS	$UB(\cdot)$	$LB(\cdot)$
\mathcal{O}_1	6.2	6.0
\mathcal{O}_2	6.0	0
\mathcal{O}_3	7.0	6.5
\mathcal{O}_4	5.0	0
\mathcal{O}_5	4.0	0

$H = \langle \mathcal{O}_3, \mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_4, \mathcal{O}_5 \rangle$,
 $\mathcal{O}_{cur} = \mathcal{O}_3, \mathcal{O}_{next} = \mathcal{O}_1$
 (d) Iteration 3

4.3 Multiple thematic relations

Holistic Top-k BD (HkBD) algorithm

Given j thematic relations $R_1^{Th}, \dots, R_j^{Th}$, we can extend analogously the original k BD algorithm by defining appropriate upper and lower bound scores for **each DS**. We can easily see that the sum of the upper (resp. lower) bound scores of all **join paths** (denote as **JP**) is the upper (resp. lower) bound score of an OS, namely:

$$UB^H(O, Q) = \sum_{\mathcal{JP}_i} UB^{\mathcal{JP}_i}(O, Q),$$

$$LB^H(O, Q) = \sum_{\mathcal{JP}_i} LB^{\mathcal{JP}_i}(O, Q),$$

where \mathcal{JP}_i ranges over all thematic paths and $UB^{\mathcal{JP}_i}(\cdot)$ (resp. $LB^{\mathcal{JP}_i}(\cdot)$) is the upper (resp. lower) bound score of O .

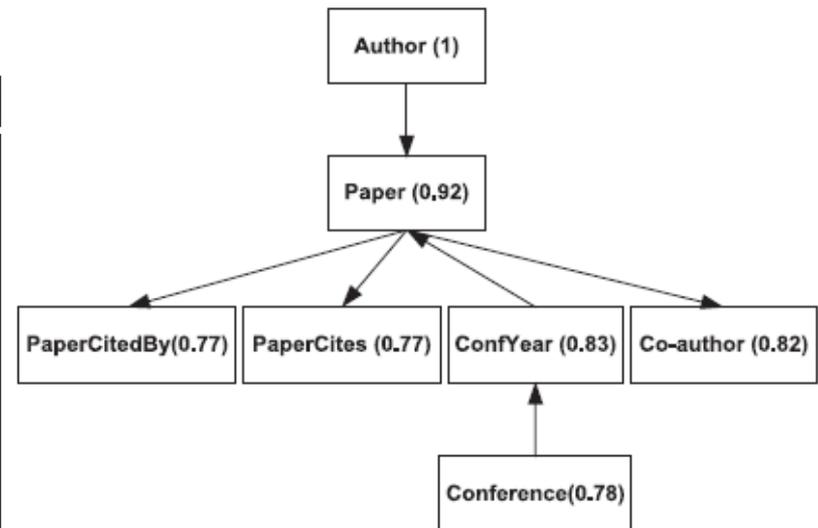
4.3 Multiple thematic relations

Holistic Top-k BD (HkBD) algorithm

We force the meeting point to the common G^{DS} prefix which is shared by all paths, then we can compute the **join result once** and reuse it later for the other paths;

HkBD approach is advantageous in this aspect over the *HBD* algorithm, as it facilitates reuse of join results.

ID	Path Name	20% DS	20% Th
\mathcal{JP}_1	DBLP: Author-Paper-Conference	68K	593
\mathcal{JP}_2	DBLP: Author-Paper-PaperCitedBy	68K	103K
\mathcal{JP}_3	DBLP: Author-Paper-PaperCites	68K	103K
\mathcal{JP}_4	TPC-H: Cust.-Ord.-Lineitem-Partsupp-Part	30K	400K
\mathcal{JP}_5	TPC-H: Cust.-Ord.-Lineitem-Partsupp-Supplier	30K	2K



Examples of Join Paths