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

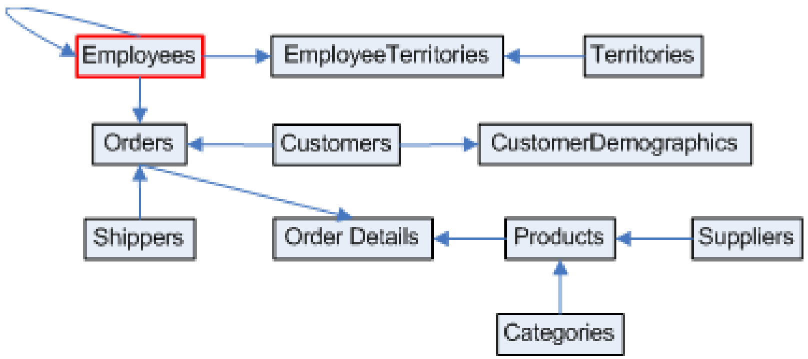
NIKOS MAMOULIS
Department of Computer Science
The University of Hong Kong
Hong Kong

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

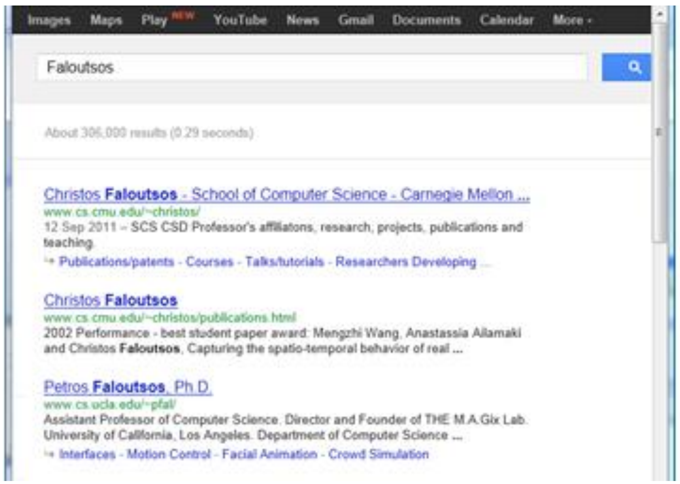
- Relational Databases are everywhere: Web, Desktops etc.
- Social graphs are also everywhere!
- Difficult to retrieve information about a **Data Subject** (DS) unless you know very well:
  - **SQL** and
  - **Schema details** etc.
- There is a need for keyword search facilities analogous to Web




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

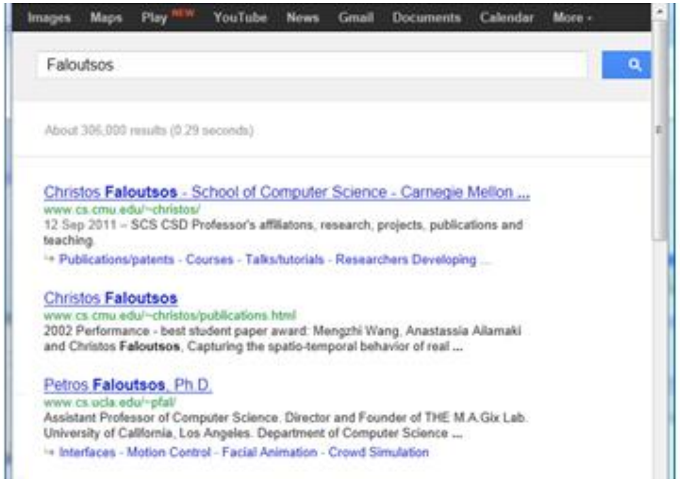
1.1 Object Summaries

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



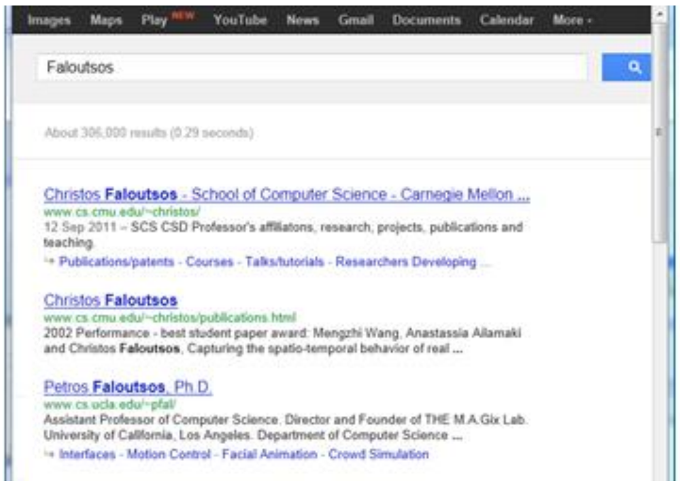
1.1 Object Summaries

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



1.1 Object Summaries

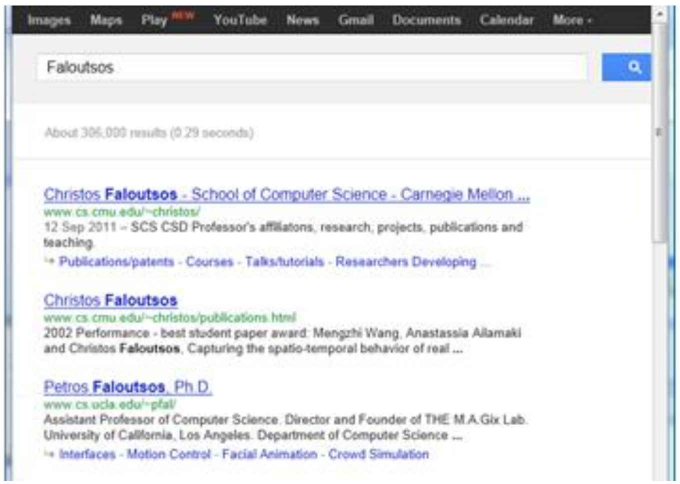
Query Search: **Faloutsos**
Web Search Result: ranked set of links and snippets



Query Search: **Faloutsos**
OS Result: set of OSs and size-*l* OSs.

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-*l* OSs.

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

	TerritoryID	TerritoryDescription	RegionID
t_1	20852	Rockville	1
t_2	27403	Greensboro	1
t_3	27511	Cary	1
t_4	30346	Atlanta	4

Region

	RegionID	RegionDescription
r_1	1	Eastern
r_2	4	Southern

Employees

	EmployeeID	LastName	FirstName	Title	TitleOfCourtesy	Address	ReportsTo
e_1	1	Davolio	Nancy	Sales Representative	Ms.	507 - 20th Ave. E. Apt. 2A	2
e_2	2	Fuller	Andrew	Vice President, Sales	Dr.	908 W. Capital Way	NULL
e_3	3	Leverling	Janet	Sales Representative	Ms.	722 Moss Bay Blvd.	2
e_4	4	Peacock	Margaret	Sales Representative	Mrs.	4110 Old Redmond Rd.	2

EmployeeTerritories

	EmployeeID	TerritoryID
et_1	3	30346
et_2	4	20852
et_3	4	27403
et_4	4	27511

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.

Orders

	OrderID	CustomerID	EmployeeID	ShipVia	ShipName	ShipAddress
o_1	10258	ERNSH	1	1	Ernst Handel	Kirchgasse 6
o_2	10273	QUICK	3	3	QUICK-Stop	Taucherstraße 10
o_3	10285	QUICK	1	2	QUICK-Stop	Taucherstraße 10
o_4	10393	SAVEA	1	3	Save-a-lot Markets	187 Suffolk Ln.
o_5	10398	QUICK	2	3	QUICK-Stop	Taucherstraße 10
o_6	10418	QUICK	4	1	QUICK-Stop	Taucherstraße 10
o_7	10451	SAVEA	4	3	Save-a-lot Markets	187 Suffolk Ln.

Shippers

	ShipperID	CompanyName	Phone
s_1	1	Speedy Express	(503) 555-9831
s_2	2	United Package	(503) 555-3199
s_3	3	Federal Shipping	(503) 555-9931

Order Details

	OrderID	ProductID	UnitPrice	Quantity	Discount
od_1	10273	2	15.2000	50	0.2
od_2	10285	2	15.2000	25	0.25
od_3	10398	2	14.4000	30	0.0
od_4	10418	1	7.6000	55	0.0
od_5	10418	2	15.2000	60	0.0
od_6	10451	2	19.2000	120	0.1

Products

	ProductID	ProductName	SupplierID	CategoryID	QuantityPerUnit	UnitsInStock
p_1	1	Chai	1	1	10 boxes x 20 bags	39
p_2	2	Chang	1	1	24 - 12 oz bottles	17

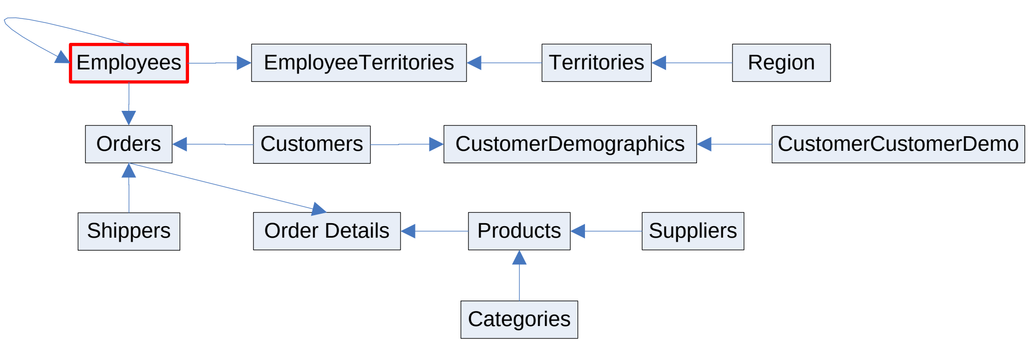
Suppliers

	SupplierID	CompanyName	ContactName	ContactTitle	Address	City
su_1	1	Exotic Liquids	Charlotte Cooper	Purchasing Manager	49 Gilbert St.	London

Categories

	CategoryID	CategoryName	Description
ca_1	1	Beverages	Soft drinks, coffees, teas, beers, ...

- 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

	TerritoryID	TerritoryDescription	RegionID
t_1	20852	Rockville	1
t_2	27403	Greensboro	1
t_3	27511	Cary	1
t_4	30346	Atlanta	4

Region

	RegionID	RegionDescription
r_1	1	Eastern
r_2	4	Southern

Employees

	EmployeeID	LastName	FirstName	Title	TitleOfCourtesy	Address	ReportsTo
e_1	1	Davolio	Nancy	Sales Representative	Ms.	507 - 20th Ave. E. Apt. 2A	2
e_2	2	Fuller	Andrew	Vice President, Sales	Dr.	908 W. Capital Way	NULL
e_3	3	Leverling	Janet	Sales Representative	Ms.	722 Moss Bay Blvd.	2
e_4	4	Peacock	Margaret	Sales Representative	Mrs.	4110 Old Redmond Rd.	2

EmployeeTerritories

	EmployeeID	TerritoryID
et_1	3	30346
et_2	4	20852
et_3	4	27403
et_4	4	27511

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.

Orders

	OrderID	CustomerID	EmployeeID	ShipVia	ShipName	ShipAddress
o_1	10258	ERNSH	1	1	Ernst Handel	Kirchgasse 6
o_2	10273	QUICK	3	3	QUICK-Stop	Taucherstraße 10
o_3	10285	QUICK	1	2	QUICK-Stop	Taucherstraße 10
o_4	10393	SAVEA	1	3	Save-a-lot Markets	187 Suffolk Ln.
o_5	10398	QUICK	2	3	QUICK-Stop	Taucherstraße 10
o_6	10418	QUICK	4	1	QUICK-Stop	Taucherstraße 10
o_7	10451	SAVEA	4	3	Save-a-lot Markets	187 Suffolk Ln.

Shippers

	ShipperID	CompanyName	Phone
s_1	1	Speedy Express	(503) 555-9831
s_2	2	United Package	(503) 555-3199
s_3	3	Federal Shipping	(503) 555-9931

Order Details

	OrderID	ProductID	UnitPrice	Quantity	Discount
od_1	10273	2	15.2000	50	0.2
od_2	10285	2	15.2000	25	0.25
od_3	10398	2	14.4000	30	0.0
od_4	10418	1	7.6000	55	0.0
od_5	10418	2	15.2000	60	0.0
od_6	10451	2	19.2000	120	0.1

Products

	ProductID	ProductName	SupplierID	CategoryID	QuantityPerUnit	UnitsInStock
p_1	1	Chai	1	1	10 boxes x 20 bags	39
p_2	2	Chang	1	1	24 - 12 oz bottles	17

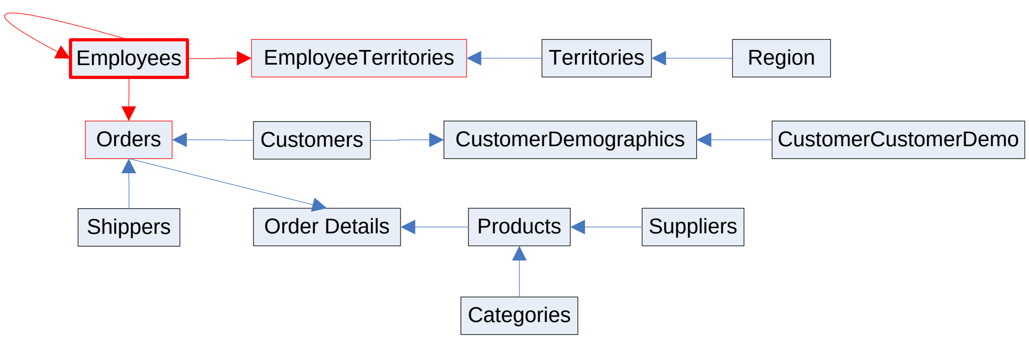
Suppliers

	SupplierID	CompanyName	ContactName	ContactTitle	Address	City
su_1	1	Exotic Liquids	Charlotte Cooper	Purchasing Manager	49 Gilbert St.	London

Categories

	CategoryID	CategoryName	Description
ca_1	1	Beverages	Soft drinks, coffees, teas, beers, ...

- 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.

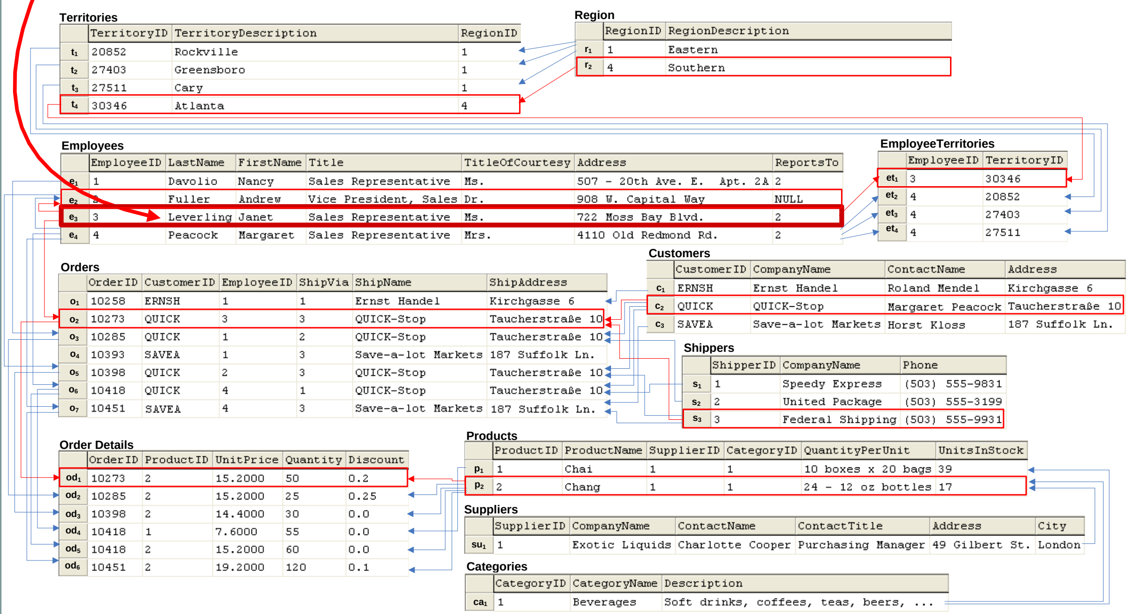


[Fakas, DKE, 2011]

2.1 Object Summaries

OS Generation - Methodology

KW-ID = “Janet Leverling”



Territories

	TerritoryID	TerritoryDescription	RegionID
t_1	20852	Rockville	1
t_2	27403	Greensboro	1
t_3	27511	Cary	1
t_4	30346	Atlanta	4

Region

	RegionID	RegionDescription
r_1	1	Eastern
r_2	4	Southern

Employees

	EmployeeID	LastName	FirstName	Title	TitleOfCourtesy	Address	ReportsTo
e_1	1	Davolio	Nancy	Sales Representative	Ms.	507 - 20th Ave. E. Apt. 2A	2
e_2	2	Fuller	Andrew	Vice President, Sales	Dr.	908 W. Capital Way	NULL
e_3	3	Leverling	Janet	Sales Representative	Ms.	722 Moss Bay Blvd.	2
e_4	4	Peacock	Margaret	Sales Representative	Mrs.	4110 Old Redmond Rd.	2

EmployeeTerritories

	EmployeeID	TerritoryID
et_1	3	30346
et_2	4	20852
et_3	4	27403
et_4	4	27511

Orders

	OrderID	CustomerID	EmployeeID	ShipVia	ShipName	ShipAddress
o_1	10258	ERNSH	1	1	Ernst Handel	Kirchgasse 6
o_2	10273	QUICK	3	3	QUICK-Stop	Taucherstraße 10
o_3	10285	QUICK	1	2	QUICK-Stop	Taucherstraße 10
o_4	10393	SAVEA	1	3	Save-a-lot Markets	187 Suffolk Ln.
o_5	10398	QUICK	2	3	QUICK-Stop	Taucherstraße 10
o_6	10418	QUICK	4	1	QUICK-Stop	Taucherstraße 10
o_7	10451	SAVEA	4	3	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	1	Speedy Express	(503) 555-9831
s_2	2	United Package	(503) 555-3199
s_3	3	Federal Shipping	(503) 555-9931

Order Details

	OrderID	ProductID	UnitPrice	Quantity	Discount
od_1	10273	2	15.2000	50	0.2
od_2	10285	2	15.2000	25	0.25
od_3	10398	2	14.4000	30	0.0
od_4	10418	1	7.6000	55	0.0
od_5	10418	2	15.2000	60	0.0
od_6	10451	2	19.2000	120	0.1

Products

	ProductID	ProductName	SupplierID	CategoryID	QuantityPerUnit	UnitsInStock
p_1	1	Chai	1	1	10 boxes x 20 bags	39
p_2	2	Chang	1	1	24 - 12 oz bottles	17

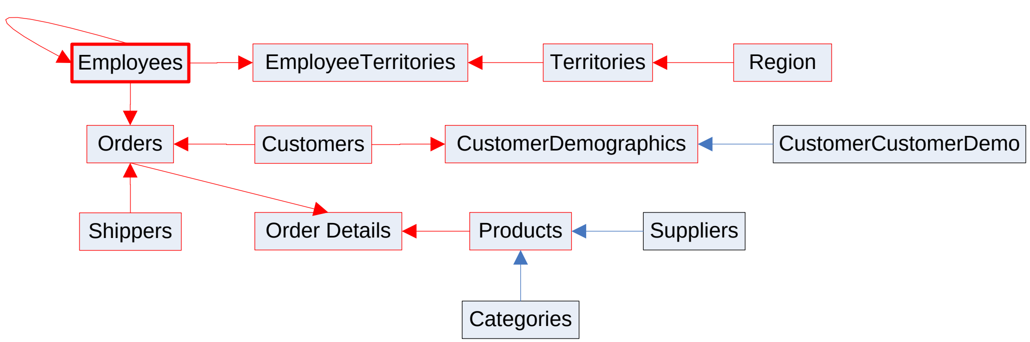
Suppliers

	SupplierID	CompanyName	ContactName	ContactTitle	Address	City
su_1	1	Exotic Liquids	Charlotte Cooper	Purchasing Manager	49 Gilbert St.	London

Categories

	CategoryID	CategoryName	Description
ca_1	1	Beverages	Soft drinks, coffees, teas, beers, ...

- 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

	TerritoryID	TerritoryDescription	RegionID
t_1	20852	Rockville	1
t_2	27403	Greensboro	1
t_3	27511	Cary	1
t_4	30346	Atlanta	4

Region

	RegionID	RegionDescription
r_1	1	Eastern
r_2	4	Southern

Employees

	EmployeeID	LastName	FirstName	Title	TitleOfCourtesy	Address	ReportsTo
e_1	1	Davolio	Nancy	Sales Representative	Ms.	507 - 20th Ave. E. Apt. 2A	2
e_2	2	Fuller	Andrew	Vice President, Sales	Dr.	908 W. Capital Way	NULL
e_3	3	Leverling	Janet	Sales Representative	Ms.	722 Moss Bay Blvd.	2
e_4	4	Peacock	Margaret	Sales Representative	Mrs.	4110 Old Redmond Rd.	2

EmployeeTerritories

	EmployeeID	TerritoryID
et_1	3	30346
et_2	4	20852
et_3	4	27403
et_4	4	27511

Orders

	OrderID	CustomerID	EmployeeID	ShipVia	ShipName	ShipAddress
o_1	10258	ERNSH	1	1	Ernst Handel	Kirchgasse 6
o_2	10273	QUICK	3	3	QUICK-Stop	Taucherstraße 10
o_3	10285	QUICK	1	2	QUICK-Stop	Taucherstraße 10
o_4	10393	SAVEA	1	3	Save-a-lot Markets	187 Suffolk Ln.
o_5	10398	QUICK	2	3	QUICK-Stop	Taucherstraße 10
o_6	10418	QUICK	4	1	QUICK-Stop	Taucherstraße 10
o_7	10451	SAVEA	4	3	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	1	Speedy Express	(503) 555-9831
s_2	2	United Package	(503) 555-3199
s_3	3	Federal Shipping	(503) 555-9931

Order Details

	OrderID	ProductID	UnitPrice	Quantity	Discount
od_1	10273	2	15.2000	50	0.2
od_2	10285	2	15.2000	25	0.25
od_3	10398	2	14.4000	30	0.0
od_4	10418	1	7.6000	55	0.0
od_5	10418	2	15.2000	60	0.0
od_6	10451	2	19.2000	120	0.1

Products

	ProductID	ProductName	SupplierID	CategoryID	QuantityPerUnit	UnitsInStock
p_1	1	Chai	1	1	10 boxes x 20 bags	39
p_2	2	Chang	1	1	24 - 12 oz bottles	17

Suppliers

	SupplierID	CompanyName	ContactName	ContactTitle	Address	City
su_1	1	Exotic Liquids	Charlotte Cooper	Purchasing Manager	49 Gilbert St.	London

Categories

	CategoryID	CategoryName	Description
ca_1	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_4, r_2)

Orders

OrderID	ShipName	ShipAddress	OrderDate	RequiredDate	ShippedDate	
10273	QUICK-Stop	Taucherstraße 10	1996-08-05	1996-09-02 0(	1996-08-12	(o_2)

Customers

CompanyName	ContactName	
QUICK-Stop	Margaret Peacock	(c_2)

Shippers

CompanyName	
Federal Shipping	(s_3)

Order Details

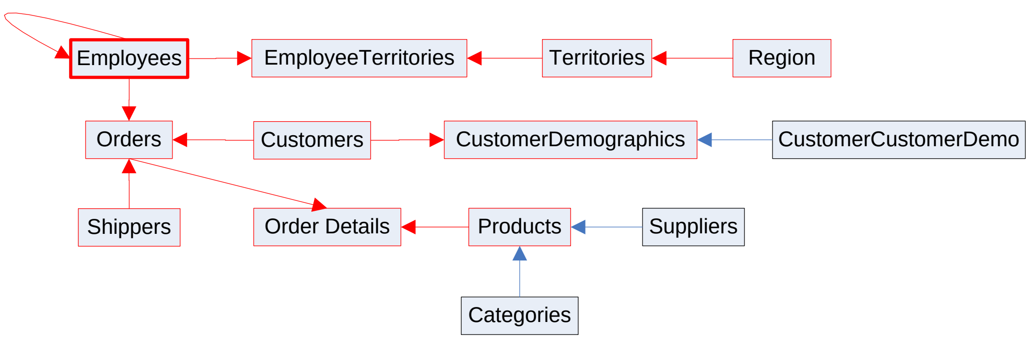
UnitPrice	Quantity	Discount	
15.2000	50	0.2	(od_1)

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' previal, 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 Themtiac *Size-l* OSs

Author: Ming-Syan <u>Chen</u> [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, $Q=<q_1, q_2>$,
- q_1 is a set of **identifying keywords**
- q_2 is a set of **thematic keywords**

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

3 Themtiac *Size-l* 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, $Q=<q_1, q_2>$,
- q_1 is a set of **identifying keywords**
- q_2 is a set of **thematic keywords**

Criterias:

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

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

where

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

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

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

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

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

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 (*k*GBJ).

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

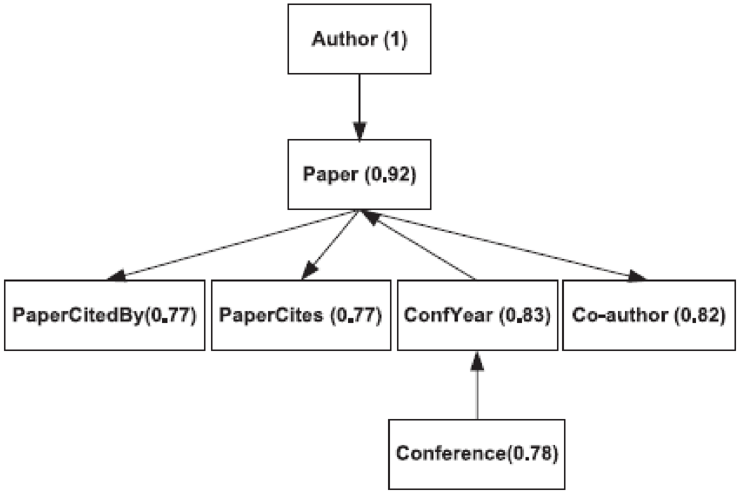


Fig. 2. DBLP Author G^{DS} (Affinity).

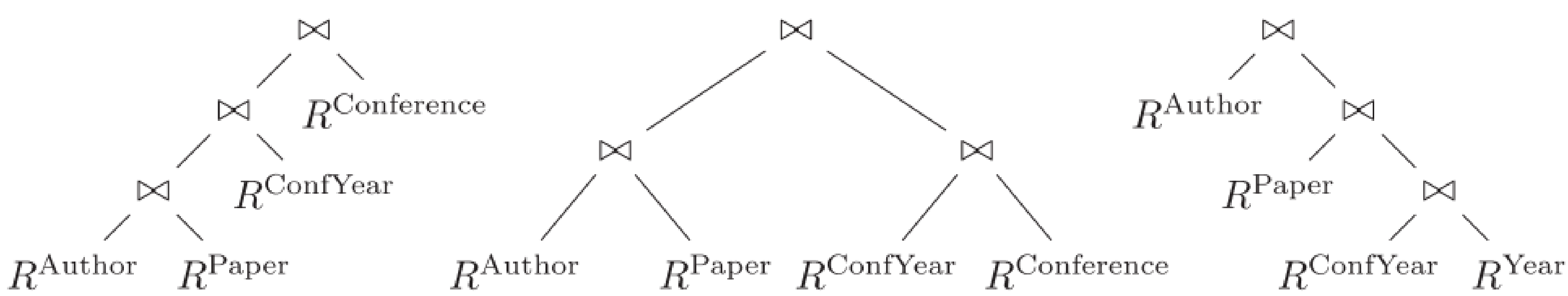
Identifying Keywords (q_1)	Frequency in DBLP
David	4,235
Chen	1,982
Wang	1,778
Alan	660
John	3,717
Nick	179

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 achieves 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

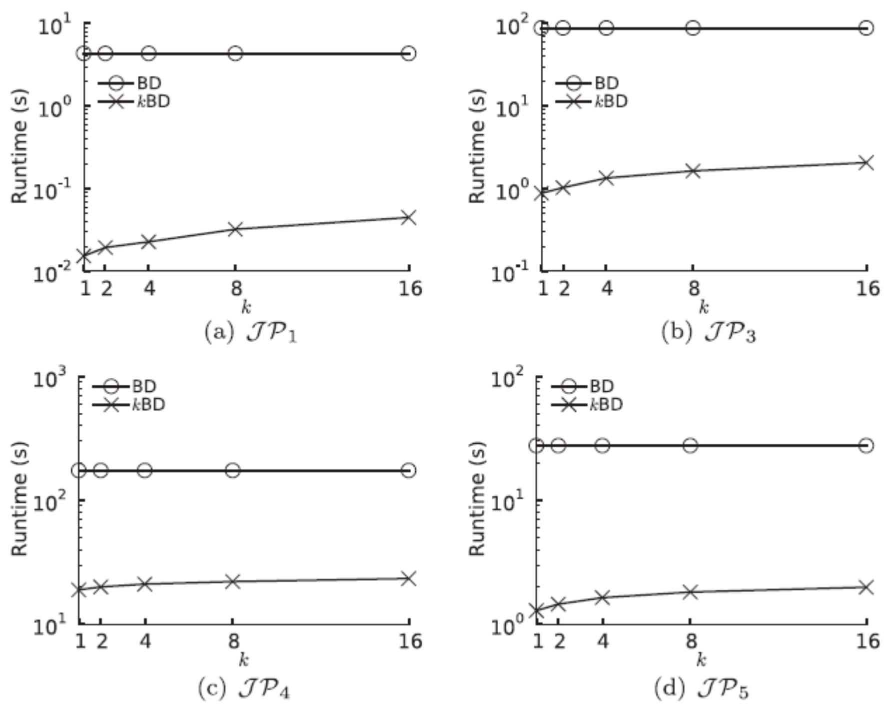
Effectiveness

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



(a) $\mathcal{JP}_1$ (b) $\mathcal{JP}_3$ (c) $\mathcal{JP}_4$ (d) $\mathcal{JP}_5$

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:

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

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

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

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

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

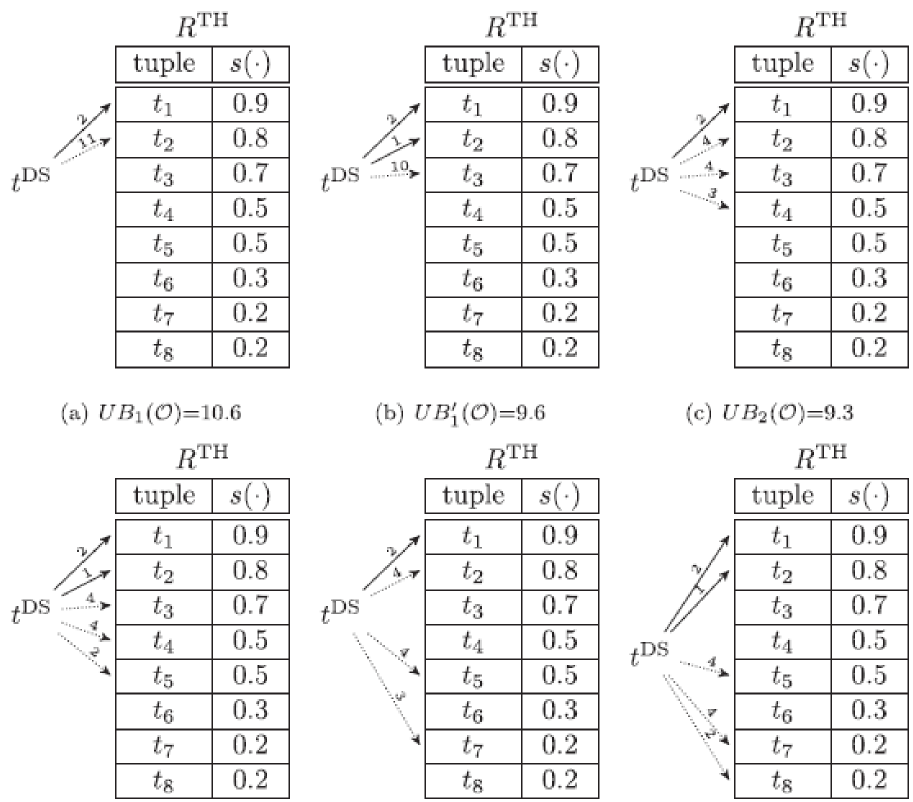
4.2 Top-*k* Bi-Directional approach

*k*BD approach

Algorithm 1. *k*BD Algorithm

kBD $(R^{DS}(q_1), R^{Th}(q_2), k)$
$H := \varnothing$;
$L^{Th} := R^{Th}(q_2)$;
sort tuples in L^{Th} in descending order of their $s(\cdot)$ scores;
for each O w.r.t. t^{DS} in $R^{DS}(q_1)$ **do**
  $LB(O) := 0$; $UB(O) :=$ CALCUB O;
  insert O into H with priority $UB(O)$;
while $k > 0 \wedge H$ is not empty **do**
  pop O_{cur} from H;
  $O_{next} := H.\text{top}()$;
  if $LB(O_{cur}) \geq UB(O_{next})$ **then**
    report O_{cur} as a result;
    $k := k - 1$;
  else
    $t_l :=$ the next tuple in L^{Th} can join with O_{cur};
    $n :=$ JOIN O_{cur}, t_l;
    $LB(O_{cur}) := LB(O_{cur}) + n \cdot In(O_{cur}) \cdot s(t_l, q_2)$;
    $UB(O_{cur}) :=$ CALCUB O_{cur};
    push O_{cur} back into H with priority $UB(O_{cur})$;

Calculating the Upper Bound Scores of an OS *O* (*In(O)* = 1.0, M = 13, m = 4)



(a) $UB_1(O)$=10.6 (b) $UB'_1(O)$=9.6 (c) $UB_2(O)$=9.3

(d) $UB'_2(O)$=8.4 (e) $UB_3(O)$=7.6 (f) $UB'_3(O)$=6.2

4.2 Top-*k* Bi-Directional approach

kBD approach

Algorithm 1. kBD Algorithm

---

kBD ($R^{DS}(q_1)$, $R^{Th}(q_2)$, k)
H:=$\varnothing$;
L^{Th}:=$R^{Th}(q_2)$;
sort tuples in L^{Th} in descending order of their $s(\cdot)$ scores;
for each O w.r.t. t^{DS} in $R^{DS}(q_1)$ **do**
  $LB(O)$:=0; $UB(O)$:= CALCUB O;
  insert O into H with priority $UB(O)$;
while $k > 0 \wedge H$ is not empty **do**
  pop O_{cur} from H;
  O_{next}:=H.top();
  if $LB(O_{cur}) \geq UB(O_{next})$ **then**
    report O_{cur} as a result;
    k:=$k - 1$;
  else
    t_l:= the next tuple in L^{Th} can join with O_{cur};
    n := JOIN O_{cur}, t_l;
    $LB(O_{cur})$:= $LB(O_{cur})$ + $n \cdot In(O_{cur}) \cdot s(t_l, q_2)$;
    $UB(O_{cur})$:= CALCUB O_{cur};
    push O_{cur} back into H with priority $UB(O_{cur})$;

The kBD Algorithm for k = 1

OS	$UB(.)$	$LB(.)$
$\mathcal{O}_1$	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	$UB(.)$	$LB(.)$
$\mathcal{O}_1$	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	$UB(.)$	$LB(.)$
$\mathcal{O}_1$	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	$UB(.)$	$LB(.)$
$\mathcal{O}_1$	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 $R_1^{Th}, \ldots, R_j^{Th}$, 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 JP_i ranges over all thematic paths and $UB^{JPi}(.)$ (resp. $UB^{JPi}(.)$) is the upper (resp. lower) bound score of *O*.

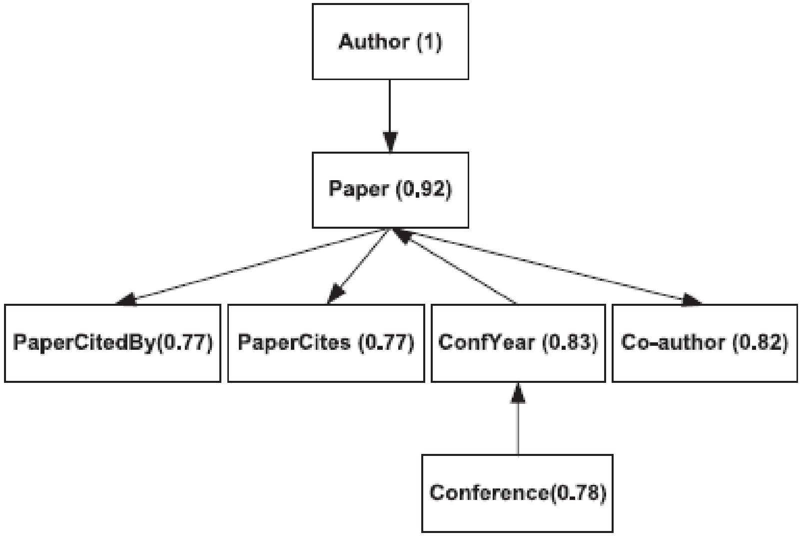
4.3 Multiple thematic relations

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 if 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-ConfYear-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