Ranked Search on Data Graphs

Ramakrishna R. Varadarajan

Doctoral Dissertation Defense

FLORIDA INTERNATIONAL UNIVERSITY,
School of Computing and Information Sciences,
Miami.

Roadmap

- Problem Statement & Motivation
- Data Model
- State of Art Graph Search Methods
- Query-Specific Summarization
- Composed Pages Search
- Explaining & Reformulating Authority-Flow Queries
- Graph Information Discovery (GID)
- Comparing Top-*k* XML Lists
- Acknowledgements

Roadmap

- **Problem Statement & Motivation**
- Data Model
- State of Art Graph Search Methods
- Query-Specific Summarization
- Composed Pages Search
- Explaining & Reformulating Authority-Flow Queries
- Graph Information Discovery (GID)
- Comparing Top-*k* XML Lists
- Acknowledgements

Problem Statement & Motivation

- Graph-structured databases – becoming a commonplace.
- Need for Efficient & High Quality search & retrieval.
- Common Graph Models:
  - *World Wide Web* (unstructured)
    - Nodes – Pages
    - Edges – Hyperlinks
  - *Relational Databases* (structured)
    - Nodes – Tuples
    - Edges – Primary/Foreign key relationships
  - *XML* (semi-structured)
    - Nodes – XML elements
    - Edges – Intra-document links (IDREFs), Inter-document links (Xlinks)

Problem Statement & Motivation

- ***Keyword Search*** – most effective & dominant information discovery method.
- Success of search engines confirm this.
- Key Advantages:
  - Simplicity (ease of use).
  - Query interface is flexible.
  - No prior knowledge about structure of underlying data.
  - Queries can be imprecise
- Recently applied over Structured (databases) & Semi-structured Data (XML).

Goals of the Dissertation

Goal - To facilitate ***user-friendly & high-quality ranked search on data graphs*** by providing solutions for:

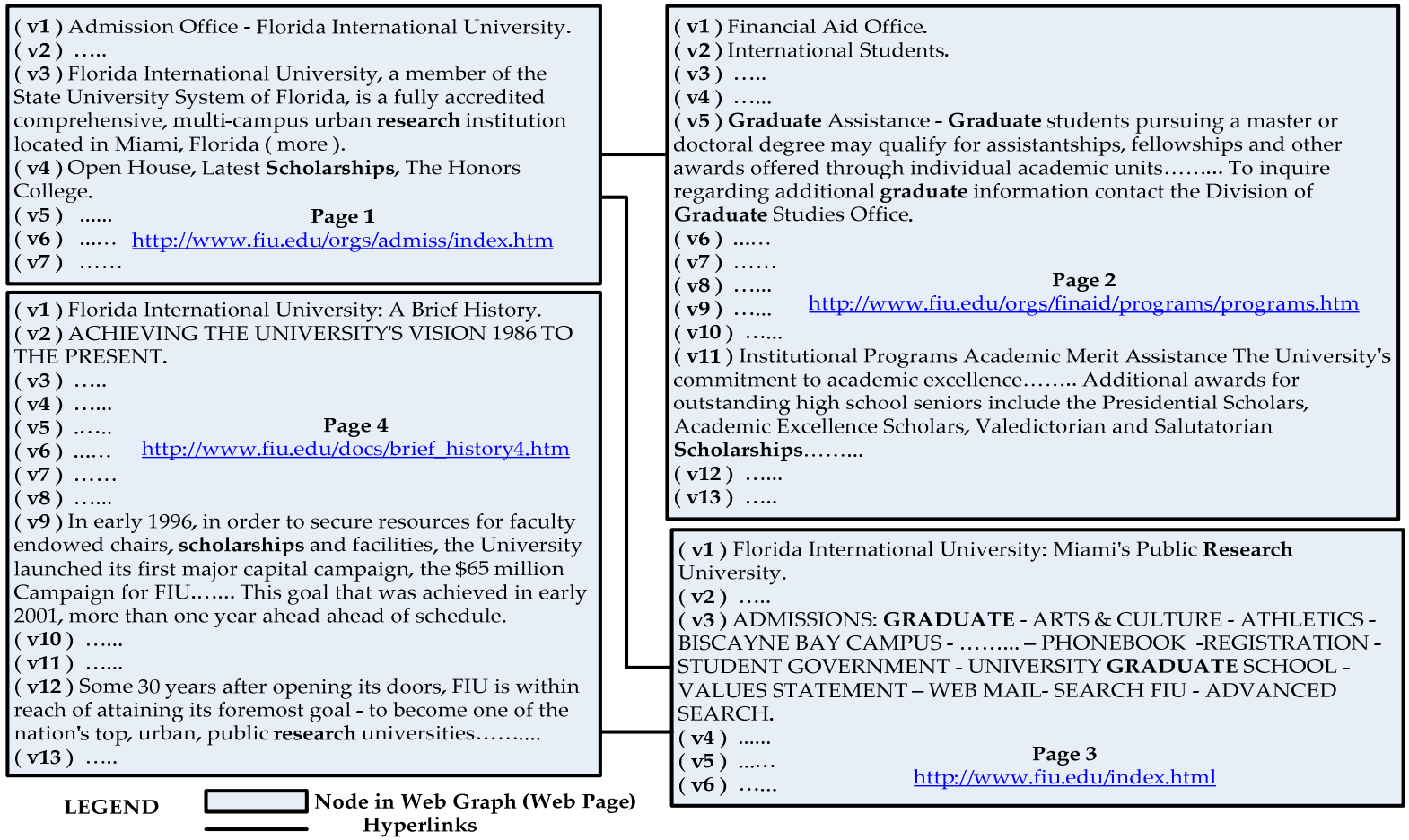
- **Result Discovery** (Composed Pages Search [SIGIR'06,TKDE'08], GID [EDBT'09], Reformulating Authority-Flow queries [ICDE'08]).
- **Result Ranking** (GID [EDBT'09], Composed Pages Search [SIGIR'06, TKDE'08], Reformulating Authority-Flow Queries [ICDE'08], Query-Specific Summarization [CIKM'05,CIKM'06]).
- **Result Presentation** (Explaining Authority-Flow Queries [ICDE'08], Query-specific Summarization [CIKM'05,CIKM'06]).
- **Evaluation of Ranked Results** (Comparing Top-k XML lists).

Roadmap

- Problem Statement & Motivation
- **Data Model**
- State of Art Graph Search Methods
- Query-Specific Summarization
- Composed Pages Search
- Explaining & Reformulating Authority-Flow Queries
- Graph Information Discovery (GID)
- Comparing Top-*k* XML Lists
- Acknowledgements

Data Model

- ***Web Graph (directed, un-weighted)***
  - ❑ Web Pages (nodes) & Hyperlinks (edges)

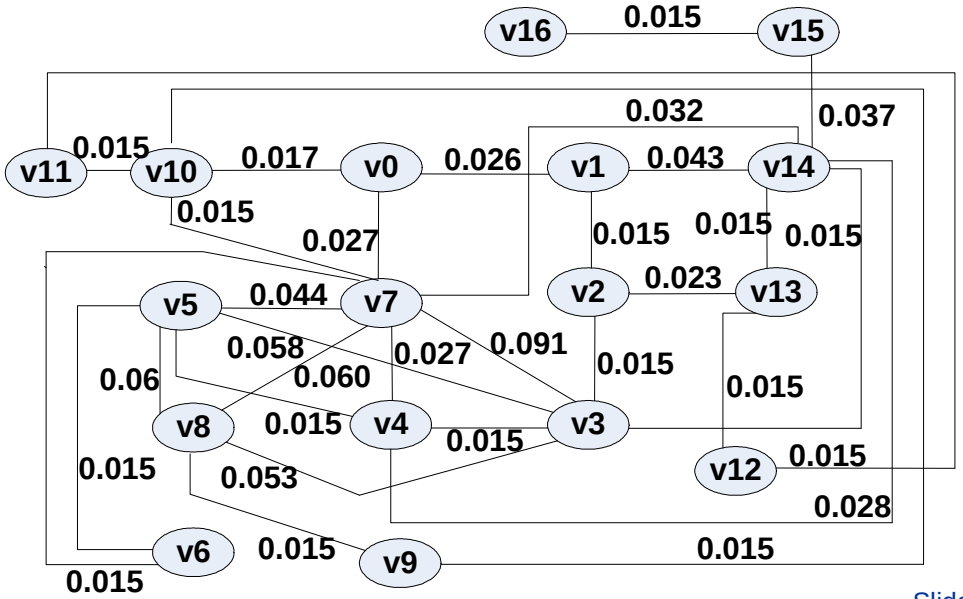


Data Model

Sample Document

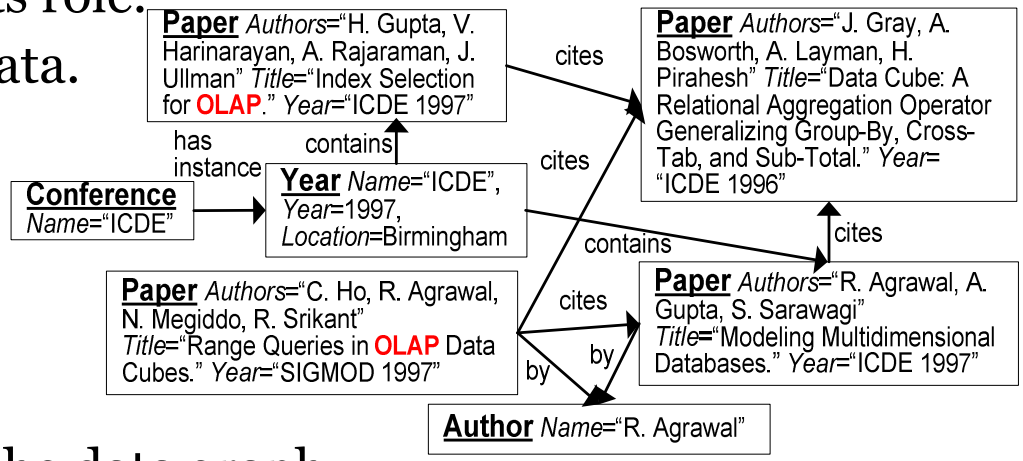
(v0) Brain chip offers hope for paralyzed
(v1) A team of neuroscientists have successfully implanted a **chip** into the **brain** of a quadriplegic man, allowing him to control a computer.
(v2) ...
(v3) ...
(v4) ...
(v5) BrainGate offers the possibility of hitherto unimaginable levels of independence for the severely disabled.
(v6) ...
(v7) ...
(v8) ...
(v9) ...
(v10) Donoghue's initial **research**, published in the science journal Nature in 2002, consisted of attaching an implant to a monkey's **brain** that enabled it to play a simple pinball computer game remotely.
(v11) The four-millimeter square **chip**, which is placed on the surface of the motor cortex area of the **brain**, contains 100 electrodes each thinner than a hair which detect neural electrical activity. The sensor is then connected to a computer via a small wire attached to a pedestal mounted on the skull.
(v12) ...
(v13)...
(v14)...
(v15) "Here we have a **research** participant who is capable of controlling his environment by thought alone -- something we have only found in science fiction so far," said Friehs.
(v16) ...

- ***Page Graph (undirected, weighted)***
  - ❖ Text Fragments (nodes)
  - ❖ Semantic links (edges)
- ✓ Parsing delimiter – NewLine.
- ✓ Text Fragments – Paragraphs.
- ✓ 17 text fragments (v0…v16).
- ✓ 17 nodes in Document Graph.

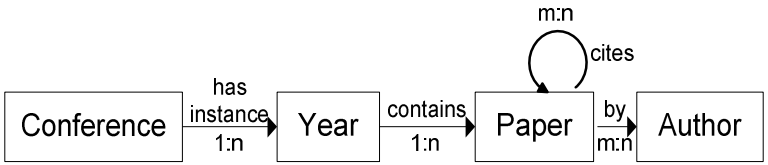


Data Model

- ***Data Graph* (directed, unweighted)**
  - ❖Tuples (nodes) & primary/foreign key relationships (edges).
  - ❖Each node represents an object & has a role.
  - ❖Each edge is labeled with its role.
  - ❖ Richer semantics – metadata.



- ***Schema Graph***
  - ❖Describes the structure of the data graph.
  - ❖High-level view of Data graph.

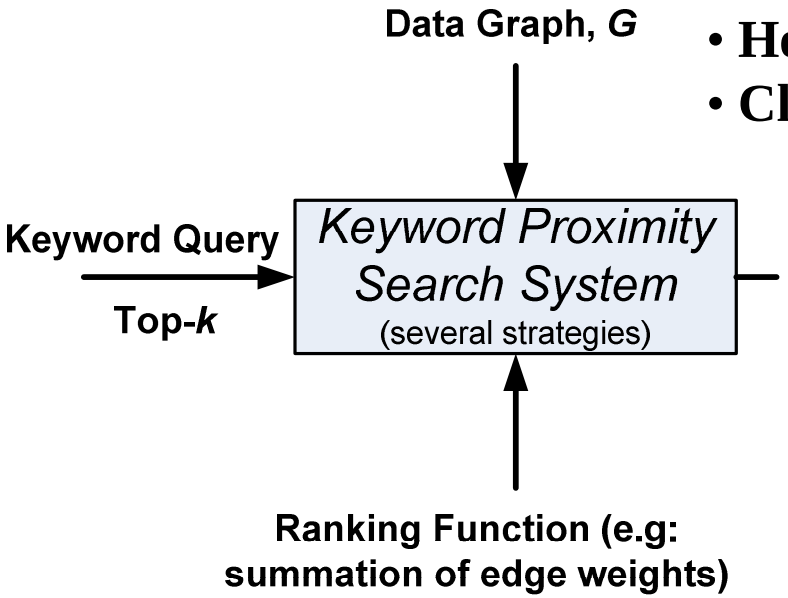


Roadmap

- Problem Statement & Motivation
- Data Model
- **State of Art Graph Search Methods**
- Query-Specific Summarization
- Composed Pages Search
- Explaining & Reformulating Authority-Flow Queries
- Graph Information Discovery (GID)
- Comparing Top-*k* XML Lists
- Acknowledgements

State of Art Graph Search Methods

- ***Keyword Proximity Search*** **(**as a black box**)**



- **How are the keywords related in the graph ?**
- **Close relationship results are ranked higher.**

Best *k* subgraphs (Spanning tree of keywords) of *G* that are
(a) Minimal (nothing more)
(b) Total (nothing less)

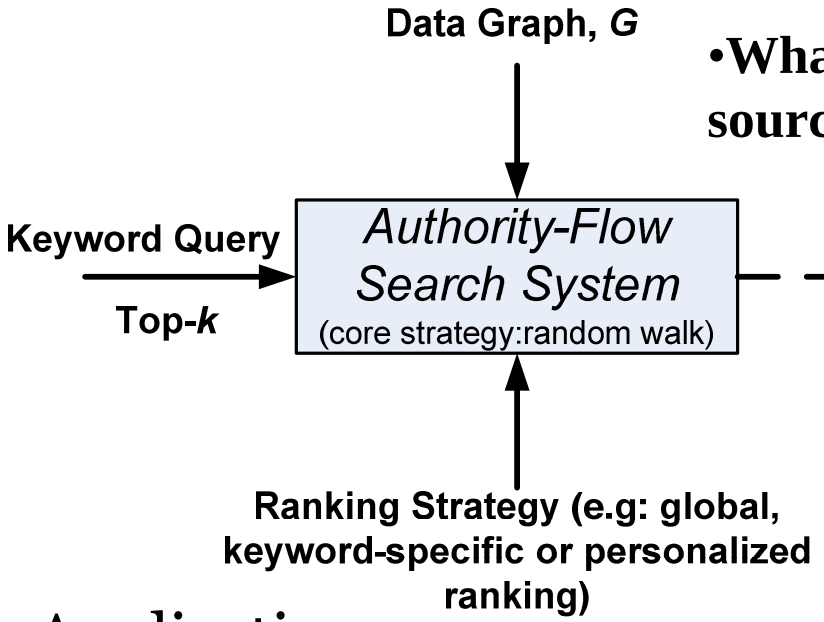
NP-Hard

Several Approximation Algorithms are proposed

- Applications:
  - **Web** ("Information Unit" paper [WWW02]).
  - **Database** (DBXplorer [ICDE02], BANKS [ICDE02], DISCOVER [VLDB02], IRStyle [VLDB03],GoldMan et. al [VLDB98]).
  - **XML** (XKeyword [ICDE03], XSearch [VLDB03]).

State of Art Graph Search Methods

- ***Authority Flow-Based Search*** **(**as a black box**)**



•What are the best *k authoritative* information sources for the query ?

Best *k* nodes of *G* of
(a) Highest Importance &
(b) Relevance to query

- Applications:
  - **Web** (PageRank [WWW98], Topic-Sensitive PageRank [WWW02], Scaling Personalized Web Search [WWW03]).
  - **Database** (ObjectRank[VLDB04]).
  - **XML** (XRANK[SIGMOD03]).

Roadmap

- Problem Statement & Motivation
- Data Model
- State of Art Graph Search Methods
- **Query-Specific Summarization**
- Composed Pages Search
- Explaining & Reformulating Authority-Flow Queries
- Graph Information Discovery (GID)
- Comparing Top-*k* XML Lists
- Acknowledgements

Summarization [CIKM'05,CIKM'06]

Motivation

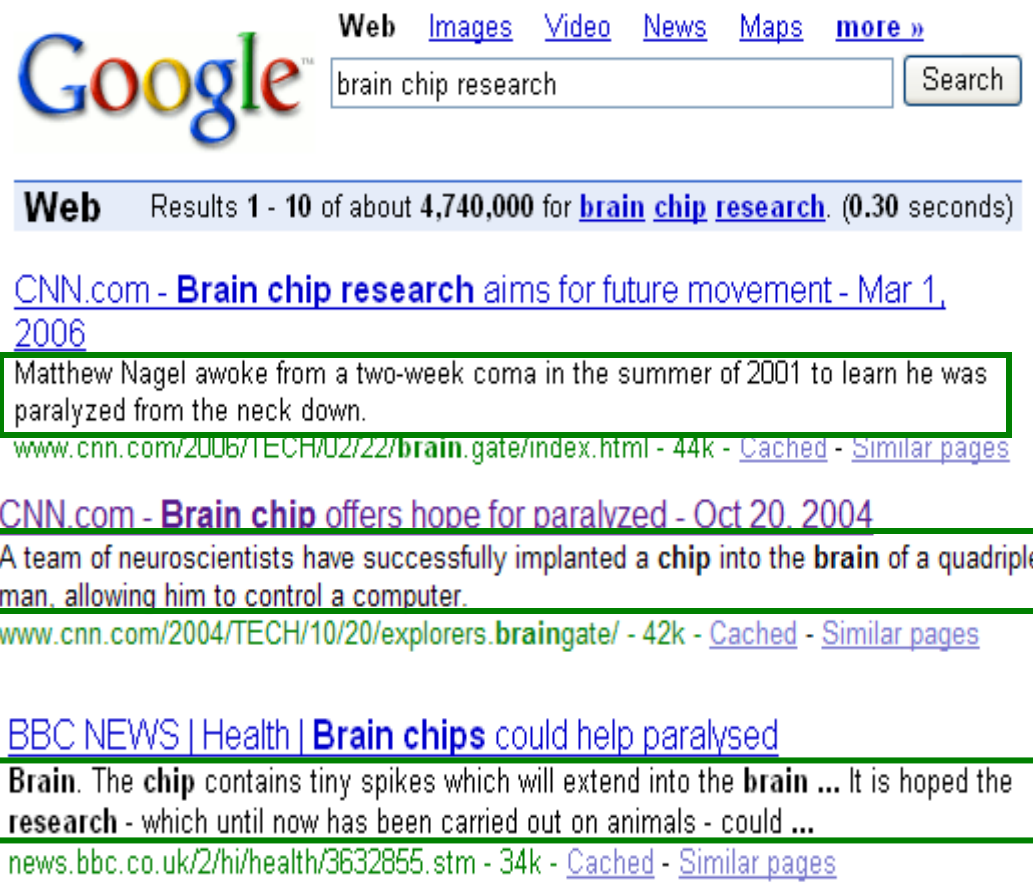
Locating relevant information on the Web is hard.

- **Summaries** are **helpful** because:
  - Provide a Quick preview of the document.
  - Allow users to quickly decide relevance.
  - Save user's browsing time.

- Two categories of summaries:
  - *Query-Independent* – Most of prior works.
  - *Query-Specific* – Applicable to web search engines.

- Success of *Web search engines* – Query specific **snippets** are important.

Summarization [CIKM'05,CIKM'06]



Motivation

Query-Specific Summaries

Summarization [CIKM'05,CIKM'06]

Drawbacks of current approaches:

- Ignores semantic relations between keywords in the document.
- Association between query keywords is unclear.
- Follows a naïve approach for query-specific summarization.

Summarization research till date:

- Mostly Query-Independent.
- Not applicable for web search.

Summarization [CIKM'05,CIKM'06]

- Document → ***graph***
- We call it ***Document Graph***.

Three Steps

Step 1: ***Preprocess***

- Build a document graph, *G*. (extract semantic relations between text fragments)

Step 2: ***Summary Generation*** (keyword proximity search)

- Given a query *Q* and a document graph *G*,

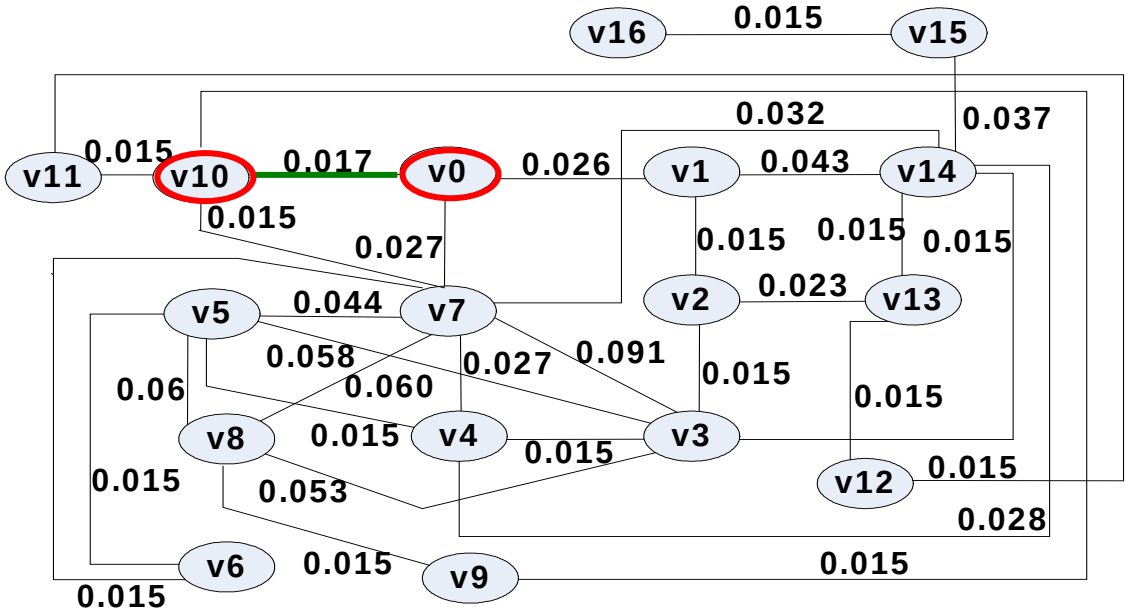
  Summaries → *Spanning Trees* that cover all keywords.

*Step 3: **Rank*** spanning trees.

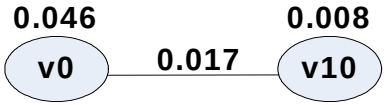
Summarization [CIKM'05,CIKM'06]

Sample Document **Example** Document Graph

(v0) Brain chip offers hope for paralyzed
(v1) A team of neuroscientists have successfully implanted a **chip** into the **brain** of a quadriplegic man, allowing him to control a computer.
(v2) ...
(v3) ...
(v4) ...
(v5) BrainGate offers the possibility of hitherto unimaginable levels of independence for the severely disabled.
(v6) ...
(v7) ...
(v8) ...
(v9) ...
(v10) Donoghue's initial **research**, published in the science journal Nature in 2002, consisted of attaching an implant to a monkey's **brain** that enabled it to play a simple pinball computer game remotely.
(v11) The four-millimeter square **chip**, which is placed on the surface of the motor cortex area of the **brain**, contains 100 electrodes each thinner than a hair which detect neural electrical activity. The sensor is then connected to a computer via a small wire attached to a pedestal mounted on the skull.
(v12) ...
(v13)...
(v14)...
(v15) "Here we have a **research** participant who is capable of controlling his environment by thought alone -- something we have only found in science fiction so far," said Friehs.
(v16) ...



Top Summary for "***Brain Chip Research***"



Score = **67.74**

Brain chip offers hope for paralyzed.
⌊ Donoghue's initial **research** published in the science journal Nature in 2002 consisted of attaching an implant to a monkey's **brain** that enabled it to play a simple pinball computer game remotely.

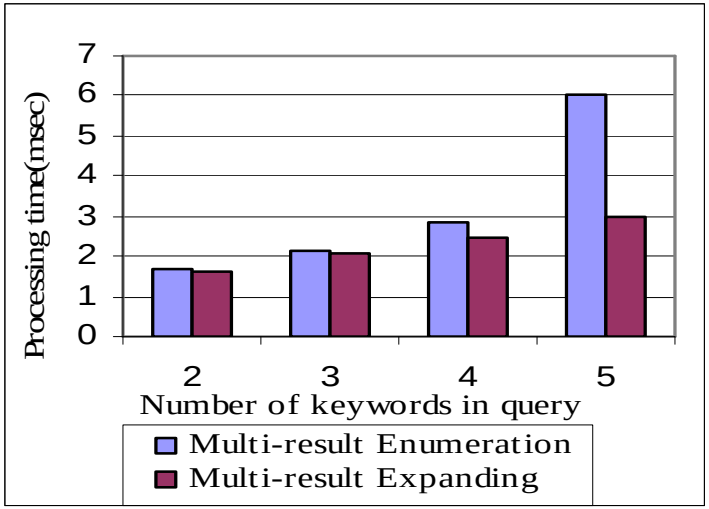
User Surveys[CIKM'05,CIKM'06]

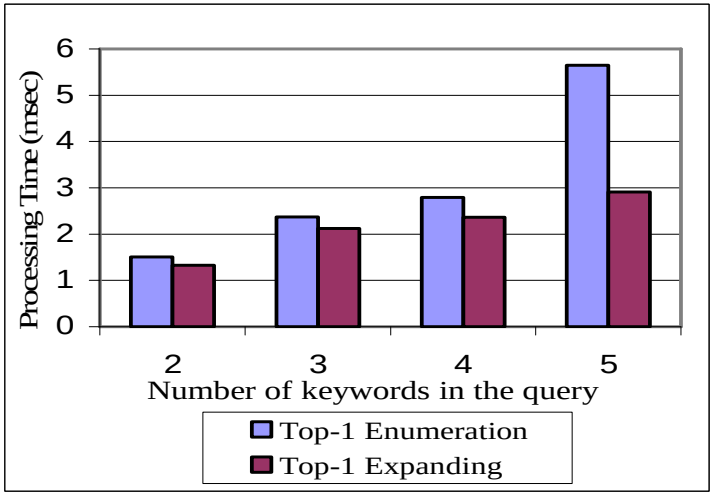
Queries	Google Desktop		MSN Desktop		Our Approach	
	D1	*D2*	*D1*	*D2*	*D1*	*D2*
1	2.33	3.67	2.33	3.67	4.87	3.67
2	2.00	3.33	2.00	3.00	4.33	3.33
3	3.00	2.67	0.67	3.00	4.93	4.00
4	1.67	2.67	1.67	3.00	4.67	4.00
5	2.00	1.67	3.00	1.00	4.00	3.67

Queries	**Document *D1***	**Document *D2***
1	Microsoft worm protection	IT Research awards
2	Anti-virus protection	Algorithms development research
3	Recovering worm deleted files	Software projects
4	Worm affected agencies	Large research grants
5	Deleted computer software	Computer network security project

Performance Experiments

News articles from ***science section*** of ***cnn.com***





Average times to calculate node weights

Number of keywords	2	3	4	5
Time (msec)	5.31	9.37	11.50	17.33

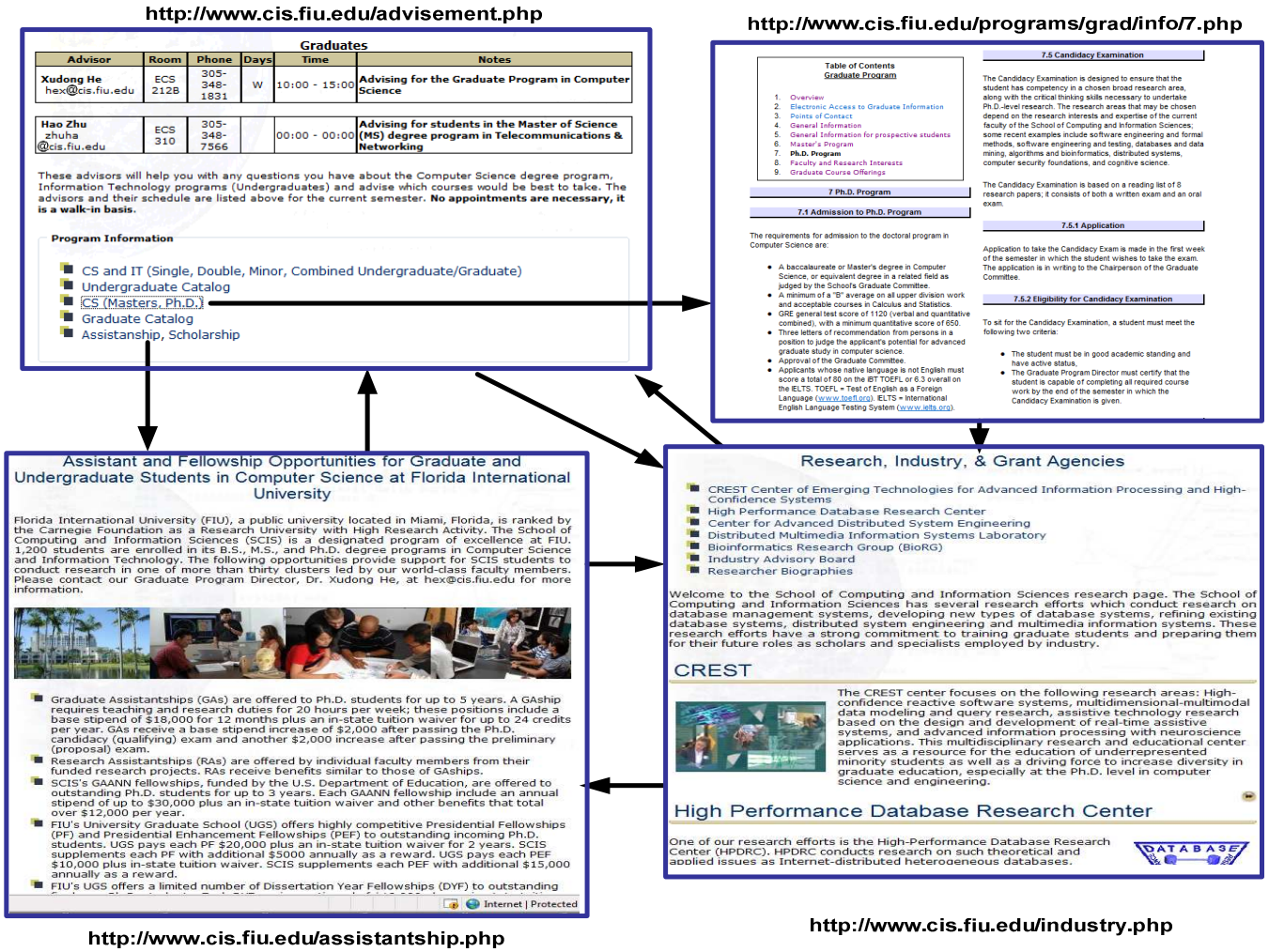
Average ranks of Top-1 Algorithms

Number of keywords	2	3	4	5
Top-1 Enumeration Algorithm.	1.4	1.8	2.1	2.78
Top-1 Expanding Search Algorithm.	1.1	1.3	1.4	1.8

Roadmap

- Problem Statement & Motivation
- Data Model
- State of Art Graph Search Methods
- Query-Specific Summarization
- **Composed Pages Search**
- Explaining & Reformulating Authority-Flow Queries
- Graph Information Discovery (GID)
- Comparing Top-*k* XML Lists
- Acknowledgements

Composed Pages Search [SIGIR'06,TKDE'08]



Motivation

Consider a *Keyword Query* –

"***Ph.D. admission requirements & fellowships***"

Each Web page only "***partially***" answers the user query.

Basic unit for search, retrieval & ranking - **individual web page**.

Current web search engines don't answer such queries completely.

Composed Pages Search [SIGIR'06,TKDE'08]

Motivation

- In WWW **- Information** is typically **distributed** across web pages & are **hyperlinked**.

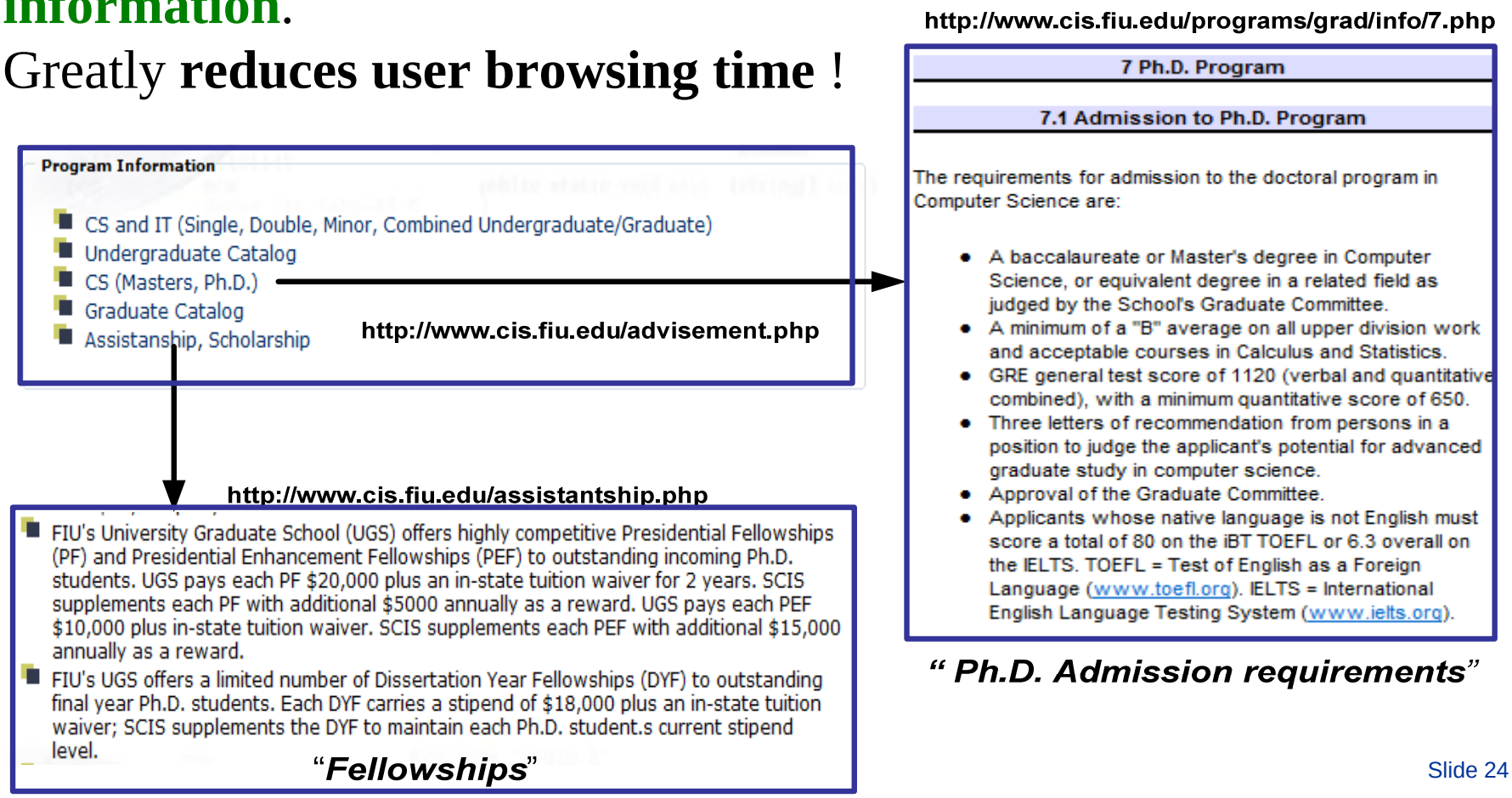
Current Web Search Engines:

- Basic unit for search & retrieval - **individual web page**.
- Return a list of ***individual*** web pages ranked by relevance.
- This degrades the quality of search results:
  - Especially for Long & Uncorrelated (multi-topic) Queries.
  - Results are not descriptive enough.
  - Does not completely satisfy users information need.
  - Users spend more time searching for relevant information.

Composed Pages Search [SIGIR'06,TKDE'08]

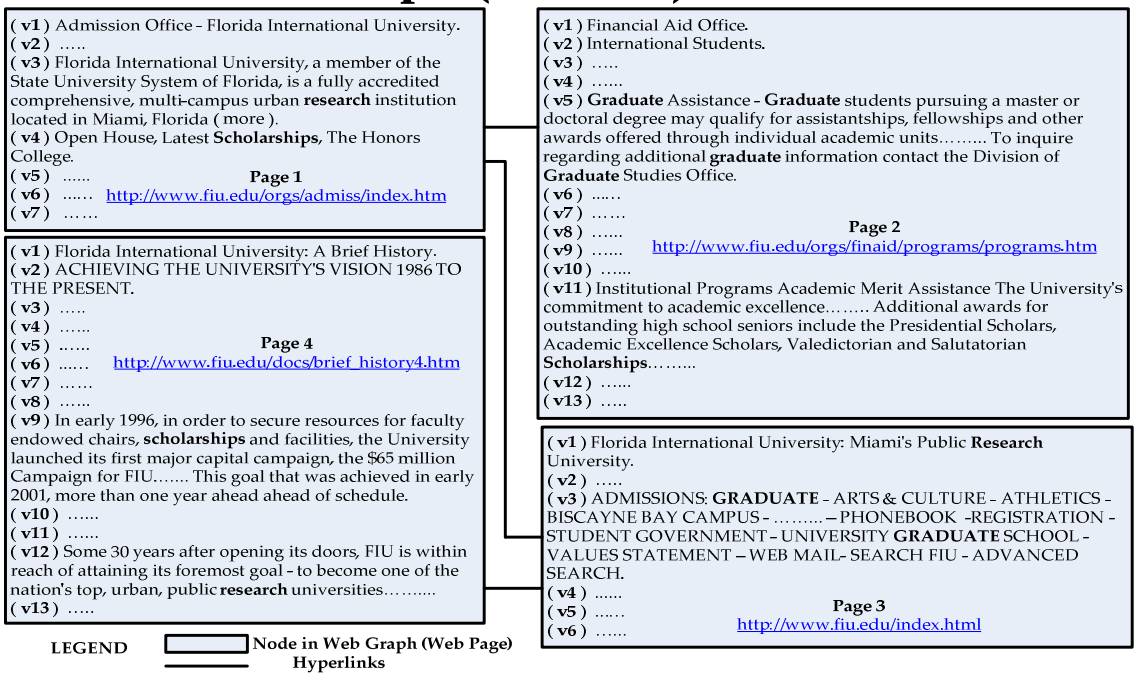
We want to extract & stitch together **"pieces"** of **relevant information**.

Greatly **reduces user browsing time** !

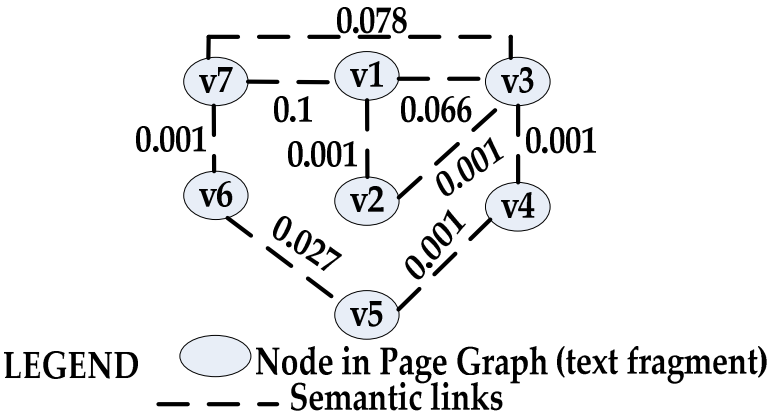


Composed Pages Search [SIGIR'06,TKDE'08]

Web Graph (crawled)



Page Graph (pre-computed)



Rank	Score	Search Results
1	12.50	1: v3 – v4 — 2: v5
2	101.60	3: v3 – v1 — 4: v9
3	209.89	3: v3 – v1 — 1: v4

We extract & stitch together pieces of information.

In contrast to previous works, we go beyond page granularity.

Composed Pages Search [SIGIR'06,TKDE'08]

Presentation & **Ranking** of Composed Pages

- First ranking principle - search results involving fewer pages are ranked higher.

$$Score\ (R) = \sum_{p \in R} \frac{Score\ (p)}{PR\ (p)}$$

- Second ranking principle - Search results of same page size, rank according to the involved page spanning trees.
- Scores of PSTs are combined using a monotone combining function.

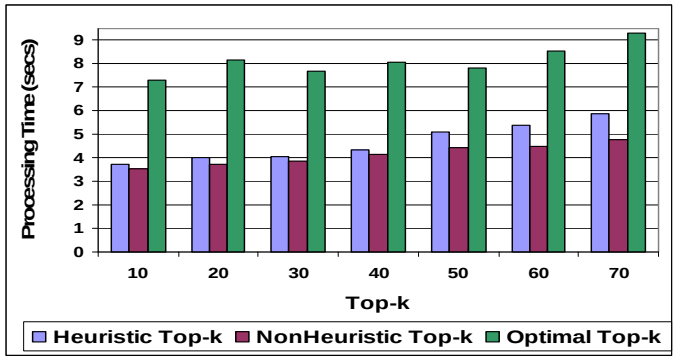
User Surveys
[SIGIR'06,TKDE'08]

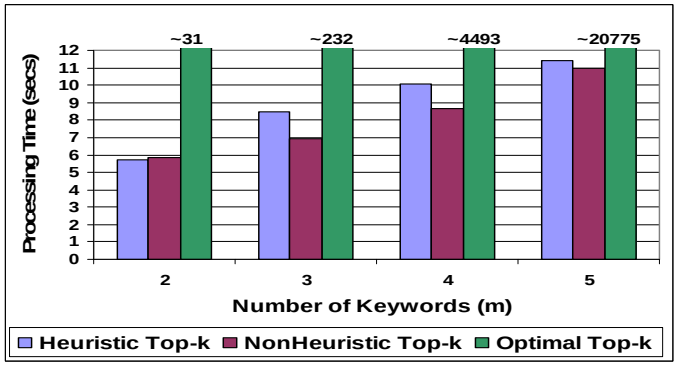
Keyword Queries	Google Search	Heuristic Expanding Search
Undergraduate Housing safety	2.06	3.41
Graduate financial aid regulations	2.41	3.59
Computer Science Internship opportunities	2.88	3.65
Campus Safety requirement regulations	2.24	3.35
Biomedical Research fellowship eligibility	1.24	3.35
Undergraduate Summer athletics accomplishments	2.25	4.5
Physics alumni achievements	3.25	3.00
Electrical transfer student eligibility	2.66	4.66
Freshman internship opportunities	1.66	4.66
Mechanical Graduate admission policies	1.66	4.66
Average Rating	**2.44**	**3.88**

Performance/Quality Experiments

Dataset: Crawled **FIU** web-pages

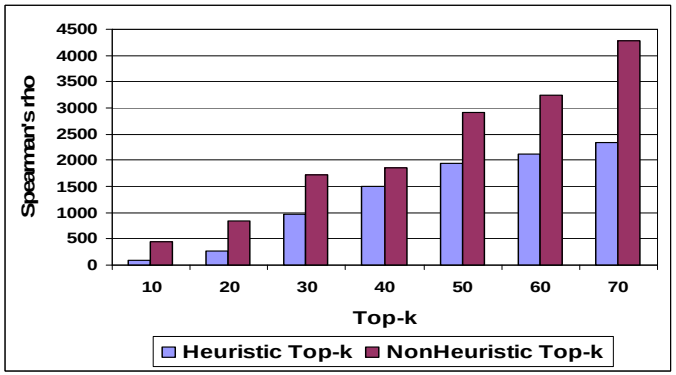
Nodes (web pages) – 25,108 & Edges (hyperlinks) - 137,929

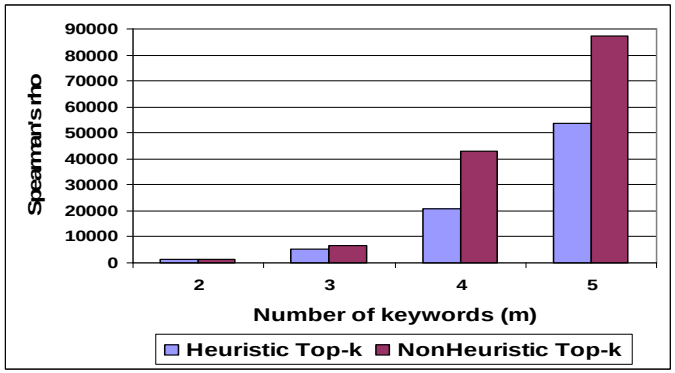




(a) Performance with changing *k* (with *m* = 2) **(b) Performance with changing *m* (with *k* = 25)**

Execution time for Top-*k* Search Results.





(a) Spearman's rho vs. Top-*k* (with *m* = 2) **(b) Spearman's rho vs. Query size (with *k* = 25)**

Quality of Algorithms.

Roadmap

- Problem Statement & Motivation
- Data Model
- State of Art Graph Search Methods
- Query-Specific Summarization
- Composed Pages Search
- **Explaining & Reformulating Authority-Flow Queries**
- Graph Information Discovery (GID)
- Comparing Top-*k* XML Lists
- Acknowledgements

Explaining & Reformulating Authority-Flow Queries [ICDE'08]

A Quick Introduction to Authority Flow Ranking:

Consider a *Bibliographic Data Graph of papers & citations* –

Simple Ranking Strategy: Papers ranked by citation count (vote).

Drawback: Each citation is given equal importance.

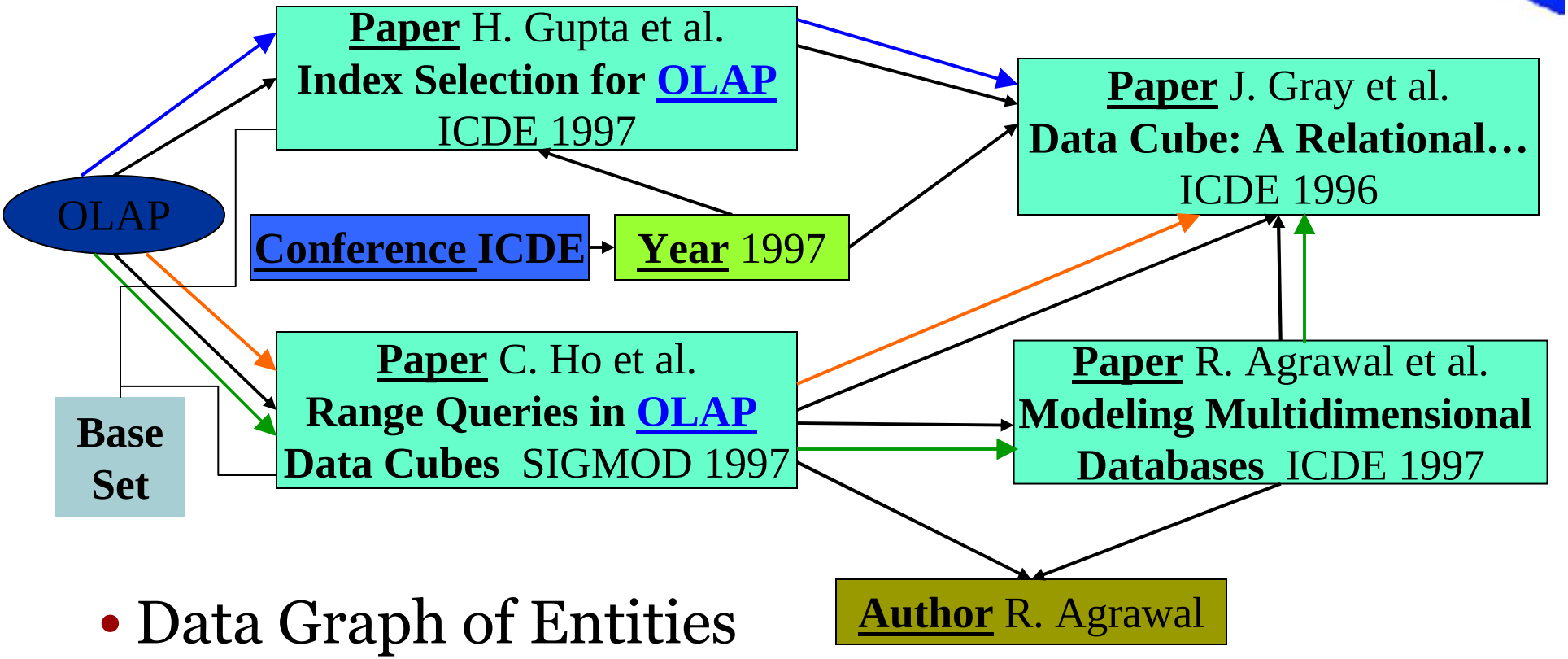
A better ranking Strategy: Papers ranked by

- Number of citations with each citation counted according to its importance.
- Importance of each citation, determined by the paper importance. (recursive in nature)
- Evenly divide the "propagated" importance to the cited papers.

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)}$$

System tunable for Global/Query-specific Importance.

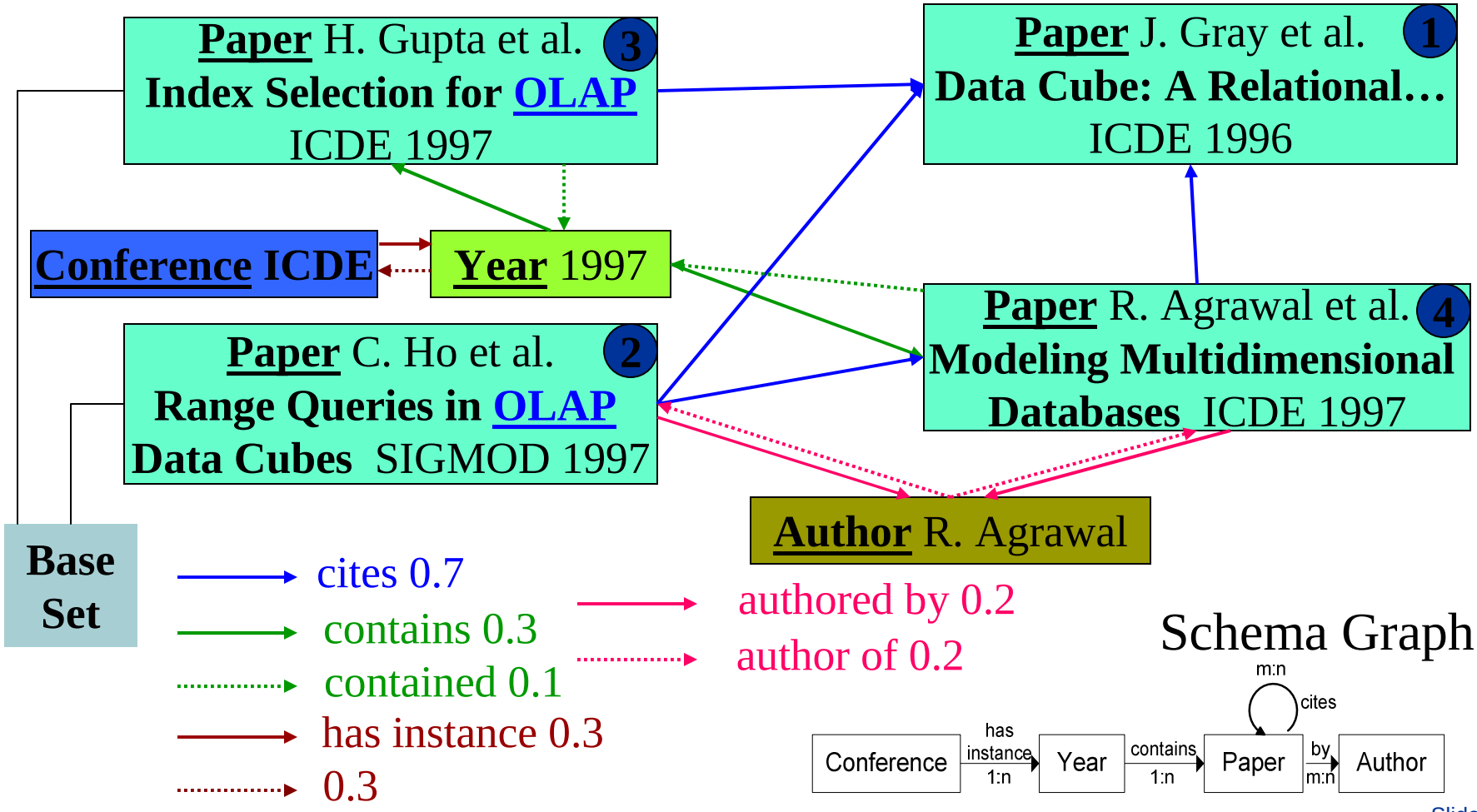
Motivation – ObjectRank [VLDB04]



- Data Graph of Entities
- **ObjectRank** Ranks Objects According to Probability of Reaching Result Starting from Base Set

Motivation – ObjectRank [VLDB'04]

Authority Transfer Data Graph (Keyword Query: [OLAP])



Slide 33

VH2

Database have edges of different types.
Different authority flows through various edges...
The authority transfer rates, which are shown at the bottom, show the maximum ratio of a node's authority transfered over edges of this type.
P->P edge has higher rate than the others because...

Another difference from the way that Web-search engines use PageRank is that we have keyword-specific ObjectRanks

Now assume we have the keyword query OLAP...

In contrast to PageRank on the Web, we can do keyword specific ObjectRanks because (a) smaller size dbs and (b) exploit schema properties to optimize algorithm.
Vagelis, 3/2/2004

Explaining & Reformulating Authority-Flow Queries [ICDE'08]

Motivation:

Top 20 results for keywords: semistructured data

1. Object Exchange Across Heterogeneous Information Sources. ICDE 1995. Hector Garcia-Molina, Jennifer Widom et al.
2. The Lorel Query Language for Semistructured Data. Int. J. on Digital Libraries 1997. Janet L. Wiener, Jennifer Widom et al.
3. Querying Semi-Structured Data. ICDT 1997. Serge Abiteboul
4. A Query Language and Optimization Techniques for Unstructured Data. SIGMOD Conference 1996. Gerd G. Hillebrand, Susan B. Davidson et al.
5. DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases. VLDB 1997. Jennifer Widom, Roy Goldman
6. Mining Association Rules between Sets of Items in Large Databases. SIGMOD Conference 1993. Arun N. Swami, Tomasz Imielinski et al.
7. Lore: A Database Management System for Semistructured Data. SIGMOD Record 1997. Dallan Quass, Roy Goldman et al.
8. Querying Semistructured Heterogeneous Information. DOOD 1995. Dallan Quass, Anand Rajaraman et al.
9. The TSIMMIS Project: Integration of Heterogeneous Information Sources. IPSJ 1994. Hector Garcia-Molina, Kelly Ireland et al.
10. Semistructured Data. PODS 1997. Peter Buneman
11. Storing Semistructured Data with STORED. SIGMOD Conference 1999. Dan Suciu, Alin Deutsch et al.
12. Adding Structure to Unstructured Data. ICDT 1997. Dan Suciu, Susan B. Davidson et al.
13. Querying Object-Oriented Databases. SIGMOD Conference 1992. Michael Kifer, Yehoshua Sagiv et al.
14. Querying the World Wide Web. PDIS 1996. Alberto O. Mendelzon, George A. Mihaila et al.
15. Mediators in the Architecture of Future Information Systems. IEEE Computer 1992. Gio Wiederhold
16. From Structured Documents to Novel Query Facilities. SIGMOD Conference 1994. Vassilis Christophides, Sophie Cluet et al.
17. Representing and Querying Changes in Semistructured Data. ICDE 1998. Serge Abiteboul, Jennifer Widom et al.
18. Relational Queries Computable in Polynomial Time Information and Control 1986. Neil Immerman
19. Extracting Schema from Semistructured Data. SIGMOD Conference 1998. Serge Abiteboul, Svetlozar Nestorov et al.
20. Programming Constructs for Unstructured Data. DBPL 1995. Dan Suciu, Susan B. Davidson et al.

Drawbacks:

- Many Top results don't contain query terms in them.
- It is ***not obvious*** why the results are relevant or important to the query.
- **Reason – ranking** primarily based on **structure** & **not content**.

Explaining & Reformulating Authority-Flow Queries [ICDE'08]

Motivation:

Limitations of Authority Flow Systems(*ObjectRank*[VLDB04]):

- No way to *explain* to the user why a particular result is relevant/important to the query.
- Authority transfer rates have to be set manually by a domain expert.
- No *query reformulation* methodology to refine results based on user-preferences.

Our Focus

- Typed domain-specific data graphs (Web search - out of scope)

Explaining & Reformulating Authority-Flow Queries [ICDE'08]

- *Problem* – Given a ***target object*** *T*, explain user why it received a high rank (or score).
- *Our Solution* – Display an **explaining sub-graph** of Authority transfer data graph, for *T*.
- Explaining sub-graph contains:
  - All Edges & corresponding Nodes that transfer authority to T.
  - Edges are annotated with amount of authority flow.
- *Steps*:
  - **Construction Stage** (using Bi-directional Breadth-First Search)
  - **Flow Adjustment Stage** (Adjust original authority flows – most challenging)

Explaining & Reformulating Authority-Flow Queries [ICDE'08]

Traditional Query Reformulation Methods:

- Well studied in Traditional IR (Salton, Buckley 1990)
- Query Expansion was the dominant strategy (ignores link-structure)
- Term selection, re-weighting, query expansion [SIGIR94, TREC95].

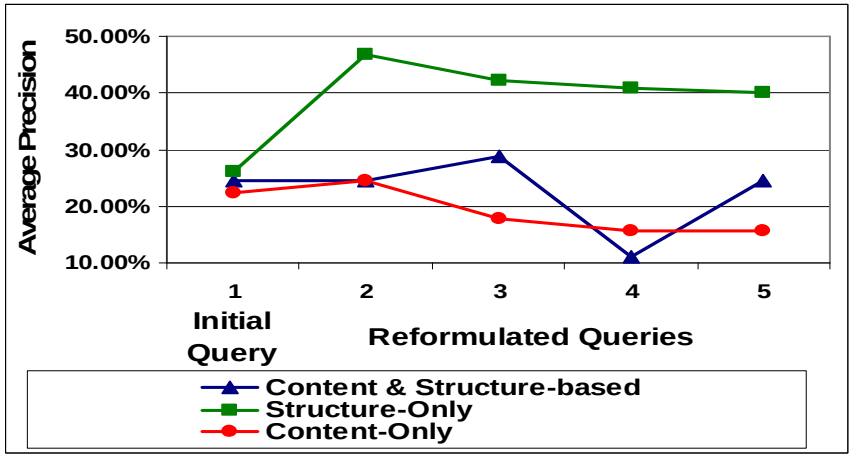
OVERVIEW OF OUR REFORMULATION ALGORITHM:

1) System computes Top-k objects with high ObjectRank2 scores.
2) User marks relevant "feedback" objects.
3) Explaining sub-graphs of feedback objects are computed.
4) Reformulate based on (a) **Content** (b) **Structure** of the graph.
5) Practically diameter is limited to a constant (L=3).

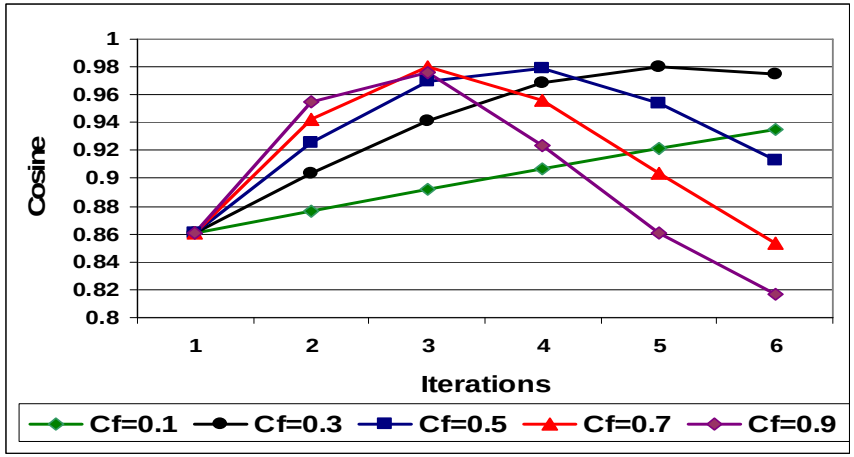
User Surveys [ICDE'08]

- Dataset: **DBLP** (Nodes - 876,110 & Edges - 4,166,626)
- Query Reformulation types tested:
  - **Content-based** Reformulations (tuning parameters - C_f=0.0 & C_e=0.2).
  - **Structure-based** Reformulations (tuning parameters - C_f=0.5 & C_e=0.0).
  - **Content & Structure-based** Reformulations (C_f=0.5 & C_e=0.2).
- 2 stages of experiments:
  - Evaluate Reformulation types (User Surveys using residual collection method).
  - Evaluate how close the trained authority transfer bounds are to the ones set by domain experts in ObjectRank [VLDB04].

(a) ***Average Precision***



(b) ***Training transfer rates***



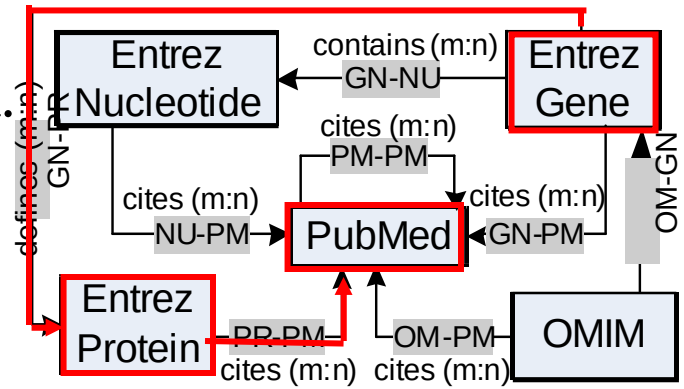
Roadmap

- Problem Statement & Motivation
- Data Model
- State of Art Graph Search Methods
- Query-Specific Summarization
- Composed Pages Search
- Explaining & Reformulating Authority-Flow Queries
- **Graph Information Discovery (GID)**
- Comparing Top-*k* XML Lists
- Acknowledgements

Graph Information Discovery (GID) [EDBT'09]

MOTIVATION

- *Consider a biologist's exploration as follows:*
  - Starting from *genes* in **Entrez Gene** she follows *links* to **Entrez Protein** and then to **PubMed**.
  - Her objects of interest are papers in **PubMed**.
  - Wants to find **PubMed papers** of highest importance/relevance to keyword "**human**".



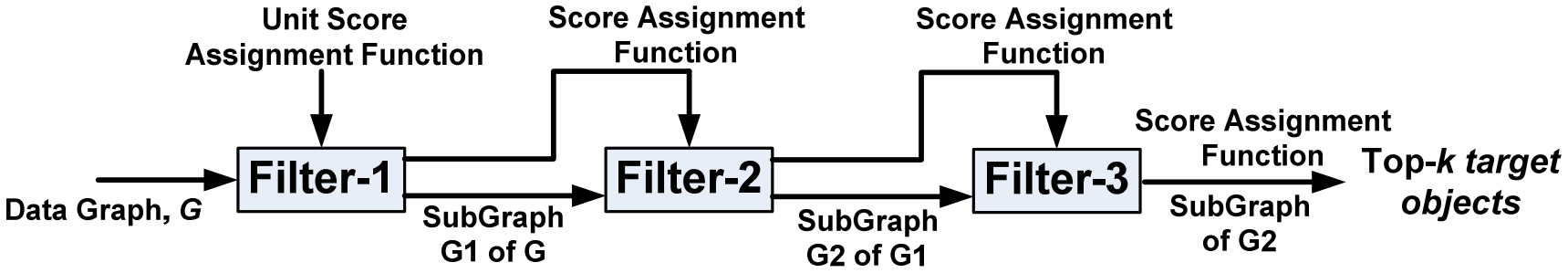
- Equivalent *graph exploration*:
  - Traverse paths **Entrez Gene → Entrez Protein → PubMed**.
  - Compute **sub-graph**.
  - Rank objects in sub-graph for query "***human***" using authority-flow.
  - **Filter** and **output** Top-*k* **PubMed Publications**.

Graph Information Discovery (GID) [EDBT'09]

- **Limitations** of current graph querying Approaches:
  - Support *extremes* of Query complexity :
    - Plain keyword queries – limited query capability.
    - Complex queries – too hard for users to learn & formulate queries.
    - Fewer solutions in between.
  - DOES NOT support:
    - *Customized* or *personalized* ranking.
    - Sophisticated *ranking techniques* like authority flow.
- **Objective**: Create a graph querying framework -
  - Easy to use & formulate Sophisticated graph queries.
  - Rank results by customized or personalized criteria.
  - Provides simple & flexible query interface.
- **Data Model**:
  - A rich web of annotated & hyperlinked data entries.
  - Includes ***schema graph*** and a ***data graph***.

Graph Information Discovery (GID) [EDBT'09]

- **GID Query Syntax & Semantics**
  - A query q is a sequence $[r_1>...>r_m]$ of FILTERS r_i.
  - A *Score assignment* function for a filter:
    - Assigns a score in [0,1] for each node.
    - Nodes with score 0 are eliminated (including its edges).



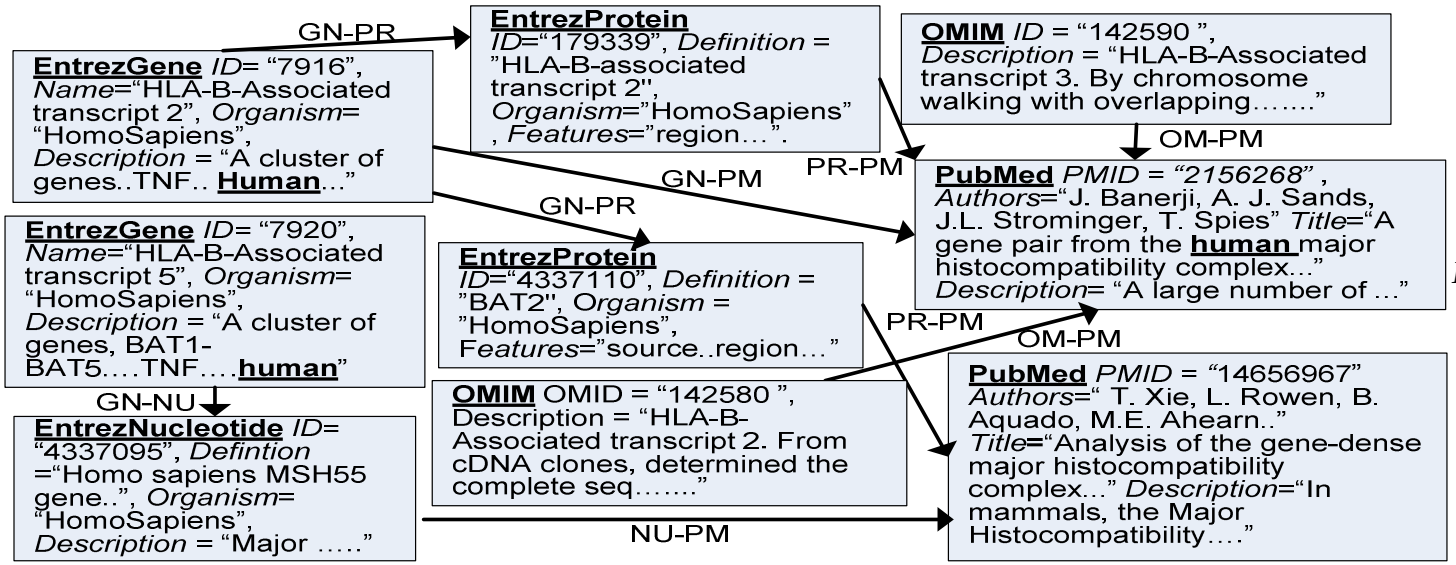
- A **filter** $r=\{R,N,S\}$ is the following 3-tuple:
  - R (Filter Selection Condition)
    - A Keyword Boolean expression (or)
    - An Attribute-value pair (or)
    - A Type (or)
    - A Path Expression
  - N (Boolean – to specify if R needs to be negated)
  - S (Boolean – Soft or a Hard Filter)

Graph Information Discovery (GID) [EDBT'09]

GID FILTER TYPES

- **HARD FILTER**
  - *Score assignment function* is ***Boolean*** (assigns score 0 or 1 to nodes).
  - Used to eliminate nodes (and their incident edges).
  - *Examples*:
    - Keyword expression *E*: Score 1 for nodes satisfying *E*.
    - Type *T*: Score 1 for nodes of type *T*; (0 otherwise).
- **SOFT FILTER**
  - Ranking is inherently fuzzy.
  - *Score assignment function* could be ***complex*** (assigns score in [0,1])
    - Authority Flow function,
    - Keyword proximity function (or)
    - IR scoring function.

Graph Information Discovery (GID) [EDBT'09]



Data Graph

Query *Q1*

HARD PATH FILTER

EntrezGene/EntrezProtein/PubMed

>

KEYWORD SOFT FILTER

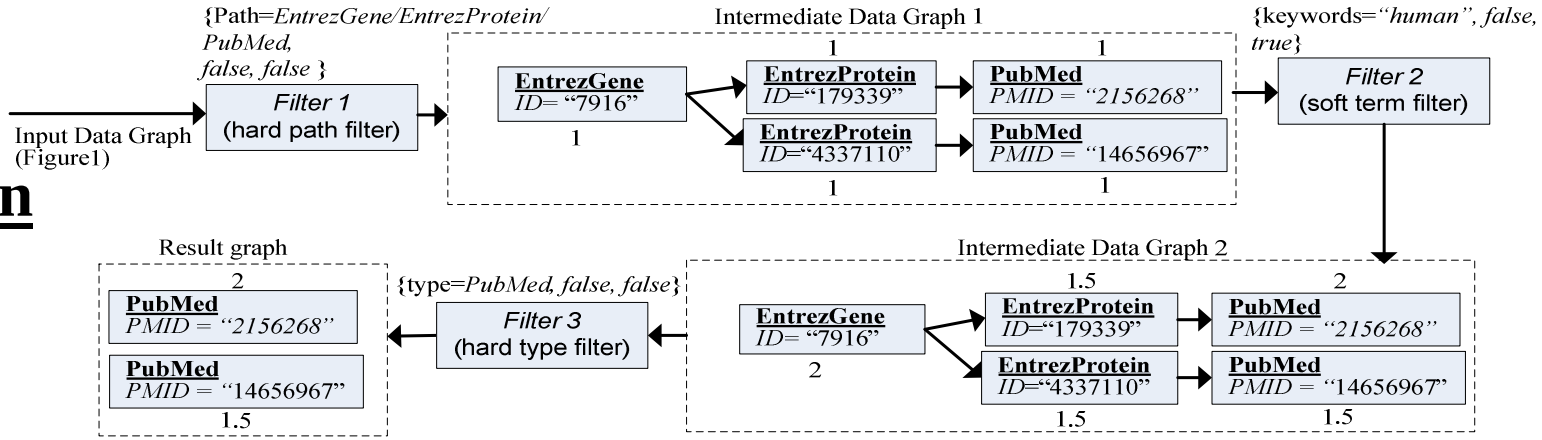
"human"

>

HARD TYPE FILTER

PubMed

Evaluation of query *Q1*

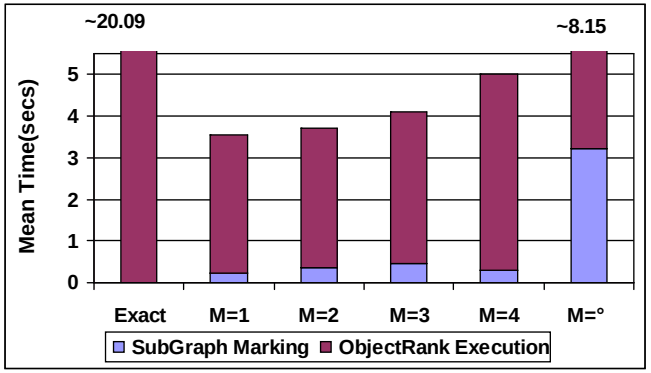


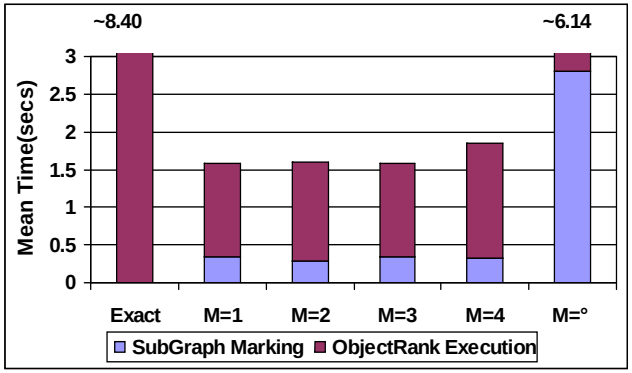
Experiments [EDBT'09]

Datasets:

- **DBLP** (Nodes - 876,110 & Edges - 4,166,626)
- **DS7** (Nodes - 699,199 & Edges - 3,533,756)

Approximating GID Soft Filters



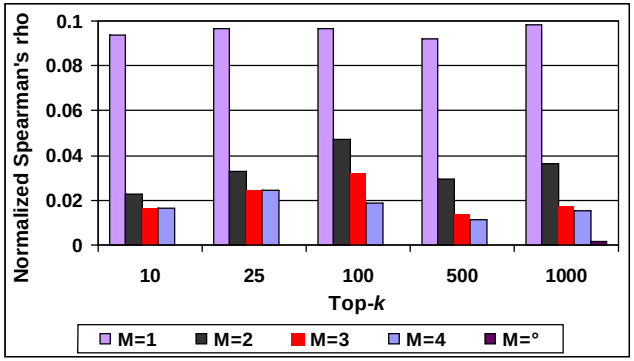


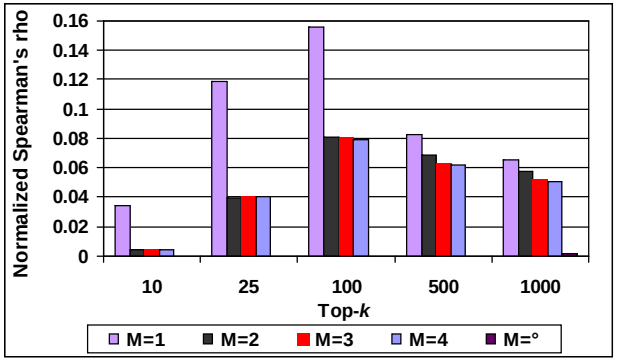
(a): DBLP Execution **(b): DS7 Execution**

Performance experiments of Path-Length-Bound Technique.

Path-Length Bound technique

Key Optimization - Limiting length of paths considered for authority flow





(a): DBLP Execution **(b): DS7 Execution**

Quality Experiments of Path-Length-Bound Technique.

Roadmap

- Problem Statement & Motivation
- Data Model
- State of Art Graph Search Methods
- Query-Specific Summarization
- Composed Pages Search
- Explaining & Reformulating Authority-Flow Queries
- Graph Information Discovery (GID)
- **Comparing Top-k XML Lists**
- Acknowledgements

Comparing Top-*k* XML Lists

- The notion of a "*top k list*" is ubiquitous in IR.
- **Objective**: Compare how similar/dissimilar two top-k Lists are based on.
  - Objects present in each list.
  - Ranking of objects.
- **Problem:** Define reasonable and meaningful distance measures between top k lists.
  - Compute a numeric distance value in [0,1]
- **Applications:**
  - Compare different search engines or variations of it.
  - Synthesize a good composite ranking function from several simpler ones (rank aggregation).
  - Design a Meta Search engine.

Comparing Top-*k* XML Lists

- **Current State-of-Art**: Distance measures for permutations and Top-k Lists [Fagin et. al SIAM'03].
  - Spearman's footrule (L1 distance)
  - Spearman's rho (L2 distance)
  - Kentall Tau
- **Objective**: Distance Measures for top-k XML Lists.

- Can we adapt existing methods ?
- Drawbacks of **existing approaches**:
  - Each object in the top-k list is viewed as a WHOLE Object.
    - **In XML – Consider 2 sub-trees differing by a single node.**
  - Matches are BOOLEAN (either match or no-match).
    - **In XML – Partial matching needed for accurate distance measures.**

Comparing Top-*k* XML Lists

BACKGROUND (Comparing individual XML trees)

- **Tree Similarity Measures:**
  - Tree-Edit, Tree-Alignment (general tree measures)
  - XML-Specific measures
    - Nierman et al. WebDB 2002 (insert-tree,delete-tree)
    - Flesca et al. WebDB 2002 (Fourier transform based Similarity)
    - Buttler 2004 (Path Shingle based Similarity)
    - Helmer VLDB 2007 (Entropy based Similarity)
    - Tag based Similarity

- **XML Lists Distance based on Total Mapping**:
  - $XLDTM(La,Lb) = a \times MinMSD^{T}(La,Lb) + b \times PD^{T}(La,Lb,fmin^{T})$

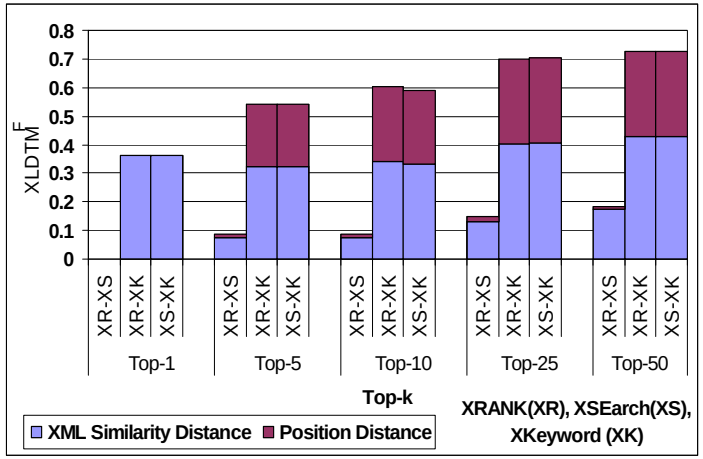
(XML Similarity Component) (Position Component)

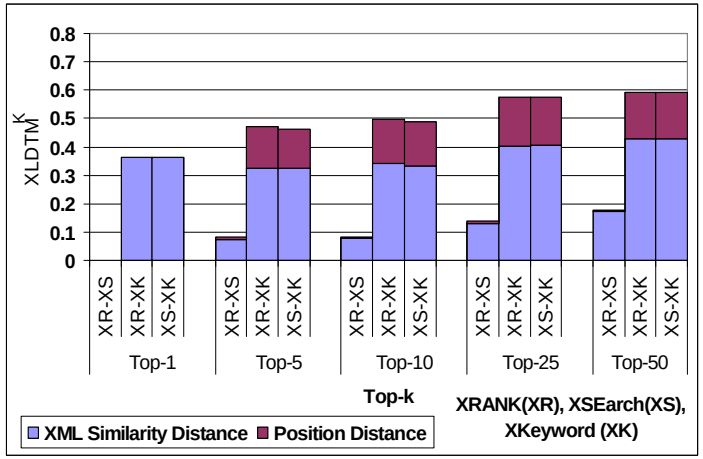
Comparing Top-*k* XML Lists

Datasets:

- **DBLP** (Elements - 7,137,933 & Average Depth - 1.90 & Max. Depth - 5)
- **NASA** (Elements - 791,923 & Average Depth - 5.58 & Max. Depth - 8)

Tree Similarity – ***Tree-Edit Distance*** **(TED)**

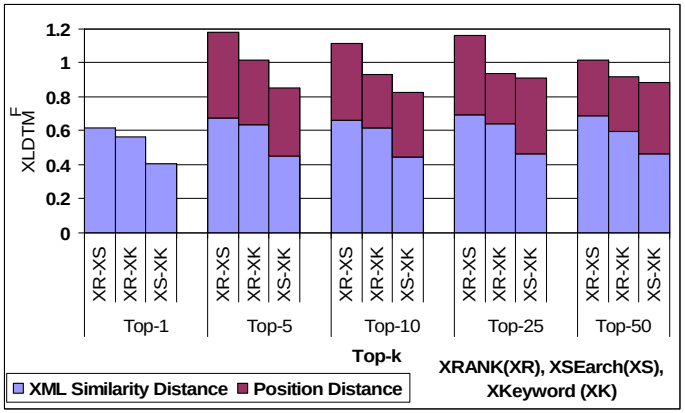


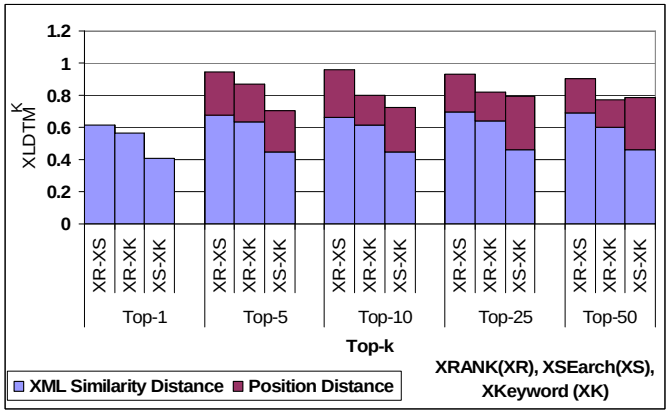


XLDTM Experiments on DBLP Dataset.

Algorithms compared:

XRANK (XR)
XSEarch (XS)
XKeyword (XK)





XLDTM Experiments on NASA Dataset.

Research (published/accepted)

- **Structure-Based Query-Specific Document Summarization**
  - ***Ramakrishna Varadarajan***, *Vagelis Hristidis*
  - Published in **ACM CIKM, 2005** (2-page poster)
- **Searching the Web Using Composed Pages**
  - ***Ramakrishna Varadarajan***, *Vagelis Hristidis, Tao Li*
  - Published in **ACM SIGIR, 2006** (2-page poster)
- **A System for Query-Specific Document Summarization**
  - ***Ramakrishna Varadarajan***, *Vagelis Hristidis*
  - Published in **ACM CIKM, 2006** (full paper)
- **Beyond Single-Page Web Search Results**
  - ***Ramakrishna Varadarajan***, *Vagelis Hristidis, Tao Li*
  - Published in **IEEE TKDE, 2008** (Journal paper)
- **Explaining and Reformulating Authority Flow Queries**
  - ***Ramakrishna Varadarajan***, *Vagelis Hristidis, Louiqa Raschid*
  - Published in **IEEE ICDE, 2008** (full paper)
- **Flexible & Efficient Querying & Ranking on Hyperlinked Data Sources**
  - ***Ramakrishna Varadarajan***, *Vagelis Hristidis, Louiqa Raschid, Maria-Esther Vidal, Luis Ibáñez, Héctor Rodríguez- Drumond*
  - Accepted for publication in **EDBT, 2009** (full paper)

Research (Current/Ongoing)

- **Comparing Top-*k* XML Lists**
  - ***Ramakrishna Varadarajan***, *Fernando Farfan, Vagelis Hristidis*
  - Under Review in **IEEE TKDE, 2009** (Journal paper)
- **Using Proximity Search to Estimate Authority Flow**
  - *Vagelis Hristidis, Yannis Papakonstantinou,* ***Ramakrishna Varadarajan***
  - Under Review in **IEEE TKDE, 2009** (Concise paper)
- **Information Discovery on Electronic Medical Records Using Authority-Flow Techniques**
  - *Vagelis Hristidis,* ***Ramakrishna Varadarajan****, Paul Biondich, Redmond Burke, Michael Weiner*
  - In preparation for **JAMIA, 2009** (journal paper)
- **Electronic Health Records**
  - *Fernando Farfan,* ***Ramakrishna Varadarajan****, Vagelis Hristidis*
  - A book chapter under review in **"Information Discovery on EHRs" book**
- **Searching Electronic Health Records**
  - ***Ramakrishna Varadarajan****, Vagelis Hristidis, Fernando Farfan*
  - A book chapter under review in **"Information Discovery on EHRs" book**
- **Web Information Extraction Using Visual Patterns**
  - ***Ramakrishna Varadarajan****, Vijil Chenthamarakshan, Prasad Deshpande, Raghuram Krishnapuram*

Roadmap

- Problem Statement & Motivation
- Data Model
- State of Art Graph Search Methods
- Query-Specific Summarization
- Composed Pages Search
- Explaining & Reformulating Authority-Flow Queries
- Graph Information Discovery (GID)
- Comparing Top-k XML Lists
- **Acknowledgements**

Acknowledgements

- **SCHOOL OF COMPUTING AND INFORMATION SCIENCES (SCIS)**
  - **Consistent Graduate assistantships (TA & RA)**
  - **Awards recognizing student research.**
  - **Conference Travel Support.**

Professor Masoud Milani

Professor Yi Deng

- **FLORIDA INTERNATIONAL UNIVERSITY**
  - **Dissertation Year Fellowship (DYF)**
  - **FIU GSA for travel awards**

Acknowledgements

Dissertation Committee (for great advise & supervision):

Professor Vagelis Hristidis
(*Academic Advisor*)

Professor Shu-Ching Chen

Professor Tao Li

Professor Raju Rangaswami

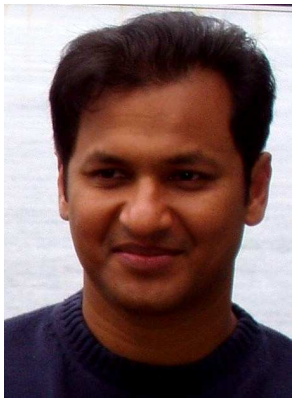
Professor Kaushik Dutta

Acknowledgements

Research Collaborators (for their support & time):

Professor Louiqa Raschid

University of Maryland at College Park



Professor Gautam Das

University of Texas at Arlington

Professor Maria-Esther Vidal

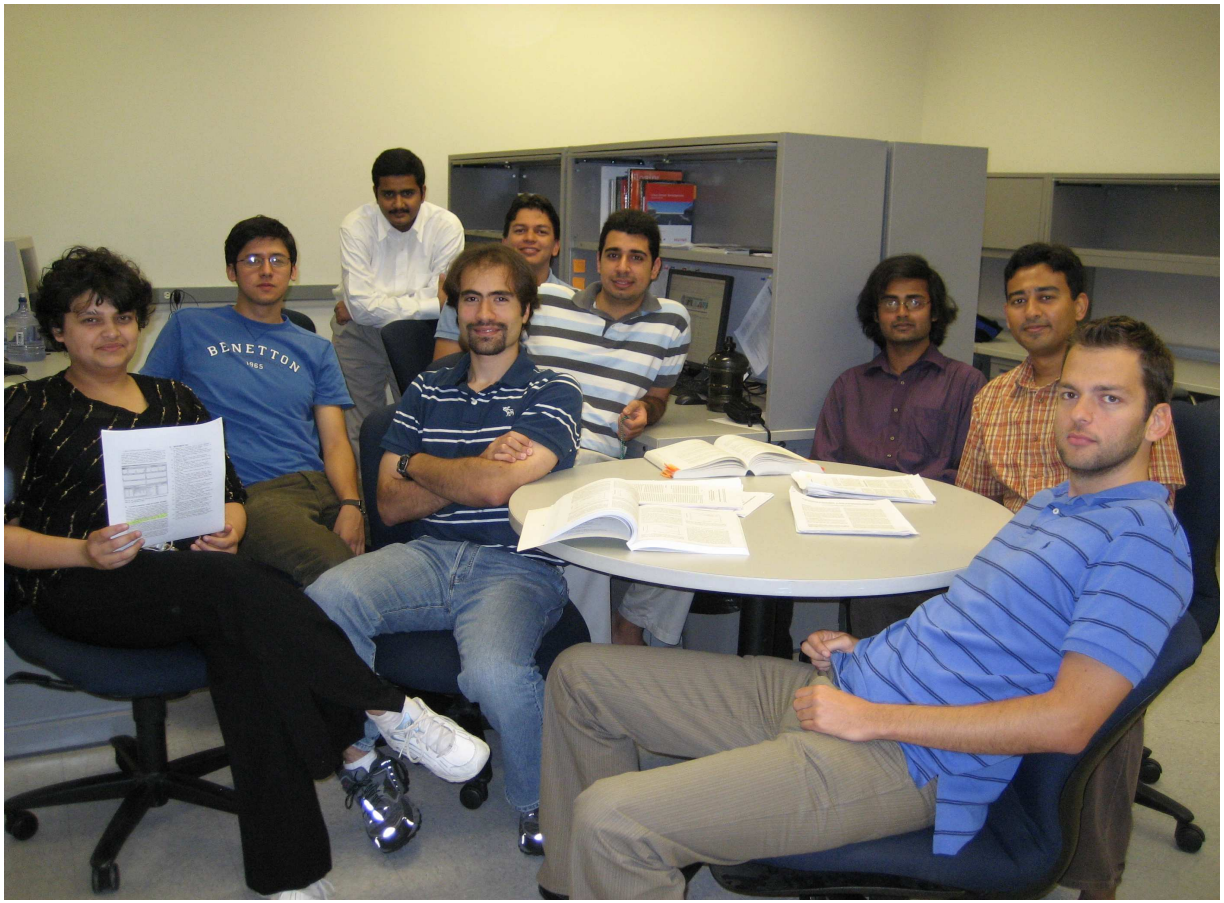
Universidad Simón Bolívar

Dr. Raghuram Krishnapuram

IBM India Research Lab

Acknowledgements

Members of Database & Systems Research Lab (DSRL):



Thanks !

Questions/Comments/

Suggestions ???

Related Work

Document Summarization

- ✓ Mostly Query-Independent
- ✓ Summarizing Web Pages
  - ❑ OCELOT - Berger et.al [SIGIR2000] synthesizes summaries (*non-extractive*).
  - ❑ INCOMMENSENSE - Paris et.al [CIKM2000] uses anchor text (ignores content).
- ✓ Splitting Web pages in to blocks
  - ❑ Song et.al [WWW2004] Block importance models (learning algorithms)
  - ❑ Cai et.al [SIGIR2004] Block level link analysis
- ✓ Document modeled as Graphs
  - ❑ Lexrank [JAIR2004] : Sentence Centrality using link analysis.
  - ❑ TextRank [EMNLP2004]: "representative" sentences using link analysis.

Keyword Search on Data Graphs

- ❖ BANKS [ICDE2002]: group-steiner tree problem
- ❖ DISCOVER [VLDB2002], DBXplorer [ICDE2002], IRStyle [VLDB2003].
- ❖ XRANK [SIGMOD2003], Xkeyword [ICDE2003] : search in XML documents.

Related Work

Information Unit [WWW Conference 2001]

Tree of hyperlinked pages containing ***ALL keywords*** - "logical **Information Unit**" (page-level)

Traditional IR Ranking

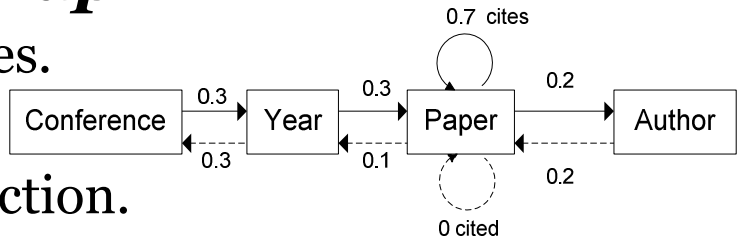
- Term weighting - State of art IR is based on tf *idf principle.
  - Okapi Formula (Modern IR overview Singhal [IEEE data bulletin 2001]).
  - Pivoted Normalized Weighting.

Link-Based Semantics

1. Web - PageRank [WWW98], HITS [ACM Journal 99], Topic-Sensitive PageRank [WWW02]
2. Database - ObjectRank for the database [VLDB02].
3. XML - XRANK [SIGMOD03].

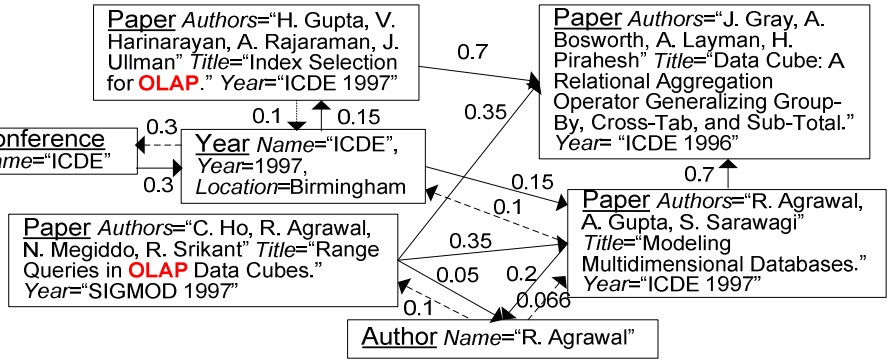
Data Model

- ***Authority Transfer Schema Graph***
  - ❖Edges reflect the authority transfer rates.
  - ❖Bi-directional authority transfer.
  - ❖Potentially different rates for each direction.



- ***Authority Transfer Data Graph* (directed, weighted)**
  - ❖Data graph edges labeled with authority transfer rates.

$$\alpha(e^f) = \begin{cases} \dfrac{\alpha(e_G^f)}{OutDeg(u, e_G^f)}, & ifOutDeg(u, e_G^f) > 0 \\ 0, & ifOutDeg(u, e_G^f) = 0 \end{cases}$$



Explaining & Reformulating Authority-Flow Queries [ICDE'08]

- **Target Object** – "*Modeling Multidimensional databases*" paper for query "*OLAP*".

Explaining Sub-graph Creation

1. Perform a BFS search in reverse direction from the target object.
2. Perform a BFS search in forward direction from base set objects (authority sources).
3. Sub-graph will contain all nodes/edges traversed in the forward direction.
4. Compute the explaining authority flow along each edge by eliminating the authority leaving the sub-graph (iterative procedure).