

Mining and Querying Multimedia Data

Fan Guo

Committee Members:

Christos Faloutsos, Chair

Eric P. Xing

William W. Cohen

Ambuj K. Singh, University of California at Santa Barbara

What this talk is about?

Carnegie Mellon Thesis Oral

COMPUTER SCIENCE DEPARTMENT

**Mining and Querying
Multimedia Data**
Fan Guo

Monday, September 19, 2011
2:00 p.m.
Class 3316

The emerging popularity of multimedia data, as digital representation of text, image, video and countless other mediums, with prodigious volumes and wild diversity, exhibits the phenomenal impact of modern technologies in reforming the way information is accessed, disseminated, digested and retained. This has iteratively ignited the data-driven perspective of research and development, to characterize perspicacious patterns, crystallize informative insights, and pattern-aptivate experience for end-users, where innovations in a spectrum of areas of computer science, including databases, distributed systems, machine learning, vision, speech and natural languages, has been incessantly absorbed and integrated to elicit the extent and efficacy of contemporary and future multimedia applications and solutions.

Under the theme of pattern mining and similarity querying, this manuscript presents a number of pieces of research concerning multimedia data, to address an array of practical tasks encompassing automatic annotation, outlier detection, community discovery, multi-modal retrieval and learning to rank, in their respective contexts including satellite image analysis, internet traffic surveillance, image bioinformatics, and Web search. A repertoire of extant and novel techniques pertaining to graph mining, clustering analysis, tensor decomposition and probabilistic graphical models has been developed or adapted, which satisfactorily met differing quality and efficiency requisites postulated by specific application settings, best exemplified by the 40 times speed-up in annotating satellite images and the up to 30% performance improvement in predicting web search user clicks, yet without the loss of generality to similar and related scenarios.

Thesis Committee:
Christos Faloutsos, Chair
Eric P. Xing
William W. Cohen
Ambuj K. Singh, University of California at Santa Barbara

5

Beyond Text and Images



6

Thesis Outline

Mining	<u>M1: MultiAspectForensics</u>
	M2: QMAS
Querying	<u>Q1: Click Models</u>
	Q2: C-DEM
	Q3: BEFH

7

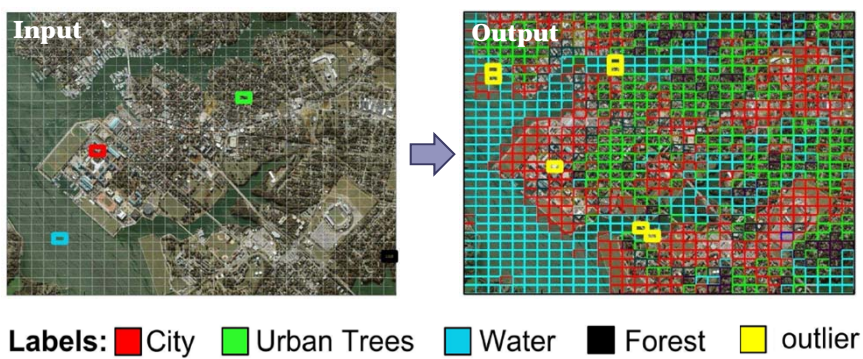
Thesis Outline

Mining	M1: MultiAspectForensics
	M2: QMAS
Querying	Q1: Click Models
	Q2: C-DEM
	Q3: BEFH

8

Mining Multimedia Data (1)

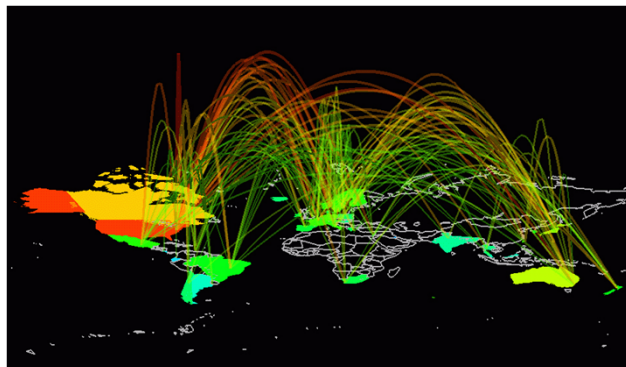
- Labeling Satellite Imagery



9

Mining Multimedia Data (2)

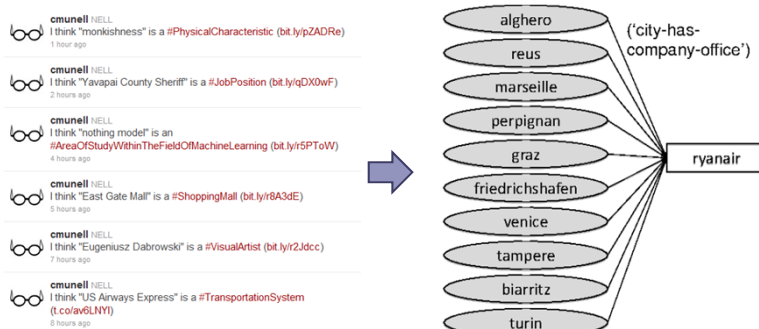
- Network Traffic Log Analysis



10

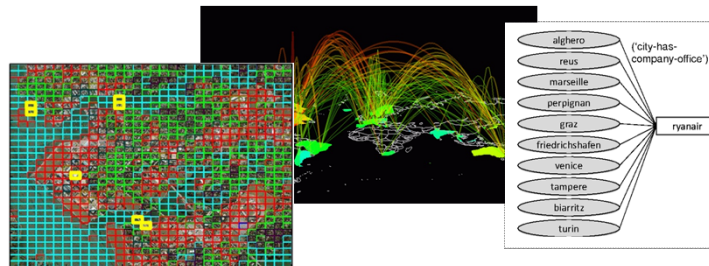
Mining Multimedia Data (3)

- Web Knowledge Base



11

Mining Multimedia Data



- Data-driven problem solving over multiple modes at a non-trivial scale.

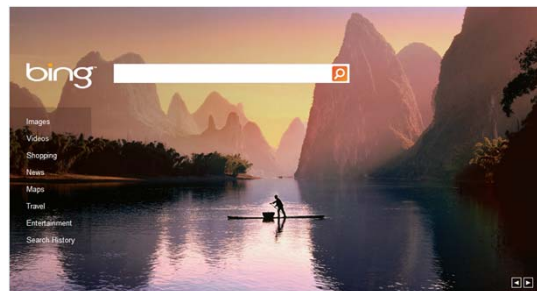
12

Thesis Outline

Mining	M1: MultiAspectForensics
	M2: QMAS
Querying	Q1: Click Models
	Q2: C-DEM
	Q3: BEFH

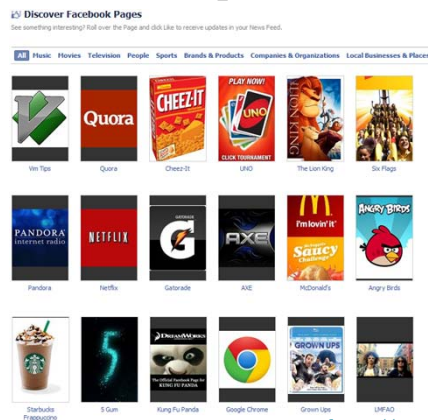
Querying Multimedia Data (1)

- A querying system provides an interface to retrieve records that best match users' information need.



Querying Multimedia Data (1)

- Here is another example:

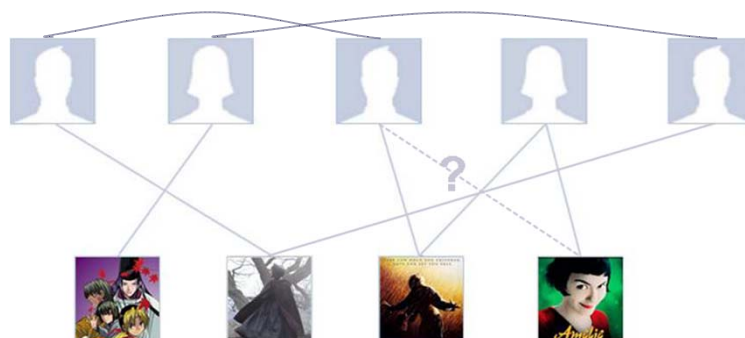


<https://www.facebook.com/pages/browser.php>

15

Querying Multimedia Data (1)

- May be transformed into a graph search problem



16

Querying Multimedia Data (2)

- Calibrate ranking from user feedback

thesis defense

About 1,840,000 results (0.13 seconds)

Advanced search

► [What is a thesis defense?](#)

www.cc.gatech.edu/faculty/ashwin/.../what-is-a-thesis-defense.html - Cached

What is a **thesis defense**? A **thesis defense** has two parts: a thesis and a defense. The second mistake many students make is not knowing what their thesis is. ...

[Thesis Defense Taboos](#)

www-psych.stanford.edu/~pinto/orals.html - Cached

148 THINGS (NOT) TO DO OR SAY AT OR FOR YOUR THESIS DEFENSE Written by Master Peter A. Dutton contributions by Jim Lalopoulos, Alison Berube, ...

[Thesis - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Thesis - Cached

In India, as in Great Britain, the **thesis defense** is called a viva voce (Latin for "by live voice") examination (viva in short). Involved in the viva are two examiners ...

[How to survive a thesis defence](#)

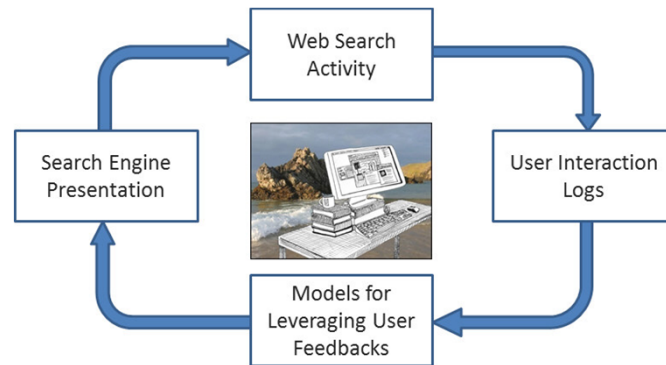
www.phys.unsw.edu.au/~jw/viva.html - Cached

This document is an appendix to: How to write a **thesis**. The **thesis defense** or viva is like an oral examination in some ways. It is different in many ways, however. ...

17

Querying Multimedia Data (2)

- Calibrate ranking from user feedback



18

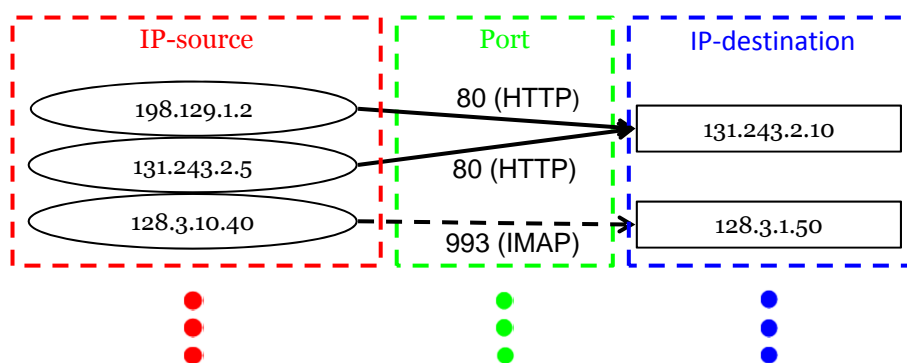
Thesis Outline

Mining	M1: MultiAspectForensics
	M2: QMAS
Querying	Q1: Click Models
	Q2: C-DEM
	Q3: BEFH

19

Data

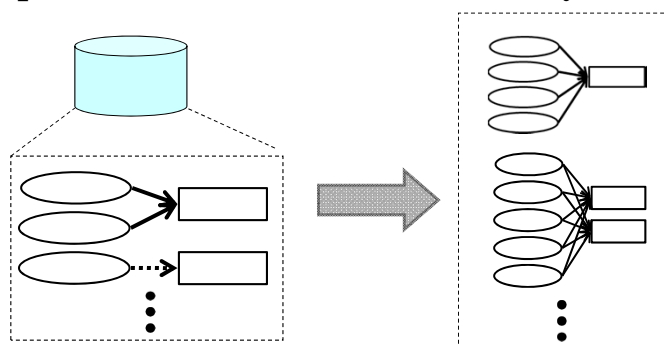
- Large-Scale Heterogeneous Networks



20

Goal

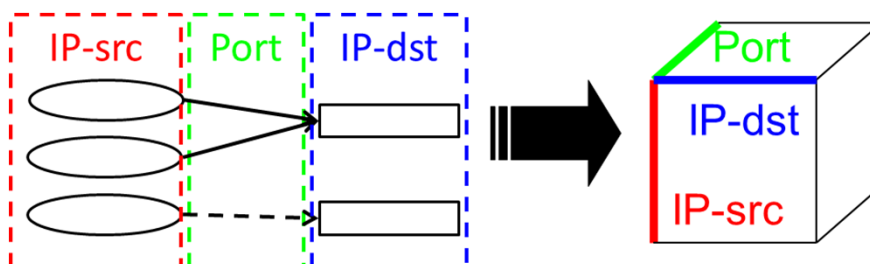
- How can we automatically detect and visualize patterns within a local community of nodes?



21

Preliminary

- Tensor for high-order data representation
 - 3 data modes: source IP, destination IP, port #



22

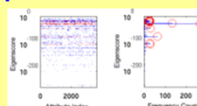
Approach

MultiAspectForensics

1. Data Decomposition

$$\mathcal{X}^{I_1 \times I_2 \times I_3} = \lambda_1 \begin{matrix} \vec{a}_1^{(1)} \\ \vec{a}_1^{(2)} \\ \vec{a}_1^{(3)} \end{matrix} + \dots + \lambda_n \begin{matrix} \vec{a}_n^{(1)} \\ \vec{a}_n^{(2)} \\ \vec{a}_n^{(3)} \end{matrix}$$

2. Spike Detection



3. Substructure Discovery

23

Data Decomposition

- The canonical polyadic (CP) decomposition can factor tensor into a sum of rank-1 tensors

$$\begin{array}{c} \textcolor{red}{I}_1 \times \textcolor{blue}{I}_2 \times \textcolor{green}{I}_3 \\ \hline \mathcal{X} \end{array} = \lambda_1 \begin{array}{|c|} \hline \overrightarrow{a}_1^{(3)} \\ \hline \overrightarrow{a}_1^{(2)} \\ \hline \overrightarrow{a}_1^{(1)} \\ \hline \end{array} + \dots + \lambda_R \begin{array}{|c|} \hline \overrightarrow{a}_R^{(3)} \\ \hline \overrightarrow{a}_R^{(2)} \\ \hline \overrightarrow{a}_R^{(1)} \\ \hline \end{array}$$

24

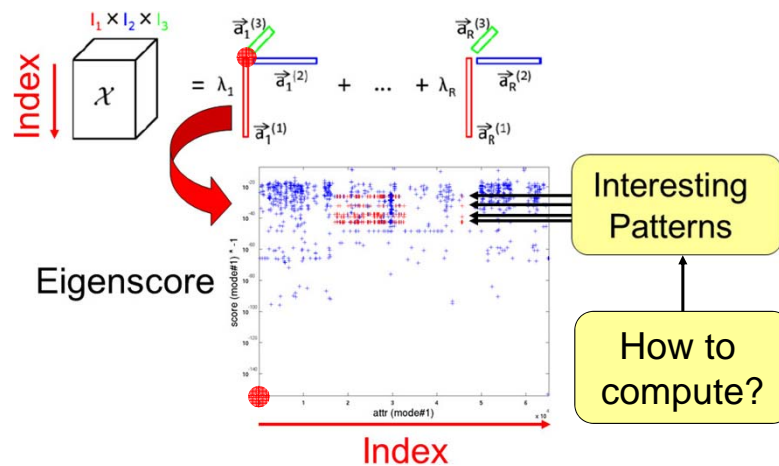
Data Decomposition

- A special case is Singular Value Decomposition

$$\begin{array}{c} \textcolor{red}{I}_1 \times \textcolor{blue}{I}_2 \\ \hline \mathcal{X} \end{array} = \lambda_1 \begin{array}{|c|} \hline \overrightarrow{a}_1^{(2)} \\ \hline \overrightarrow{a}_1^{(1)} \\ \hline \end{array} + \dots + \lambda_R \begin{array}{|c|} \hline \overrightarrow{a}_R^{(2)} \\ \hline \overrightarrow{a}_R^{(1)} \\ \hline \end{array}$$

25

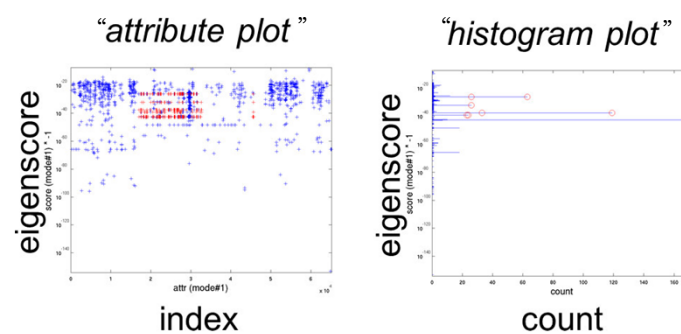
Attribute Plot



26

Spike Detection

- Iteratively search for spikes in the histogram plot along each data mode.



27

Substructure Discovery

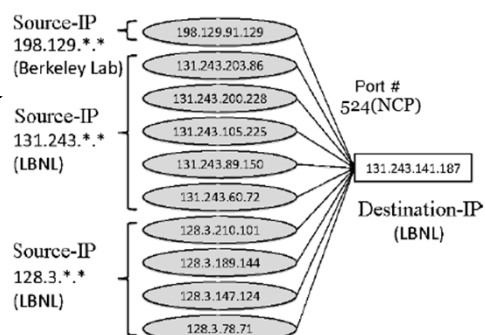
- Focus on part of the data within the spike
- Categorize into a few subgraph patterns

28

Pattern 1: Generalized Star (1)

IP-src's sending packets to the same IP-dst & the same port

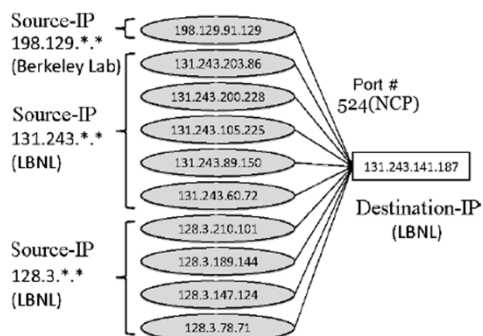
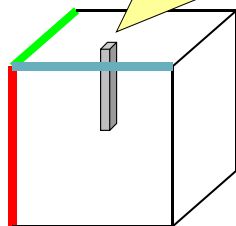
Typical client/server system



29

Pattern 1: Generalized Star (1)

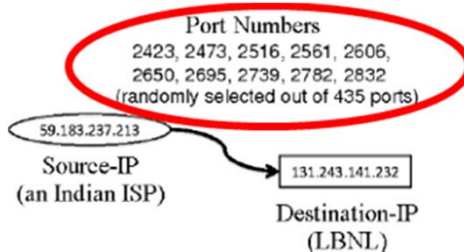
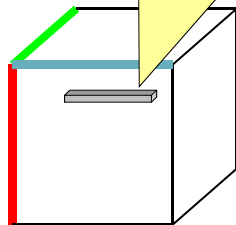
A 'bar' in a carefully reordered tensor



30

Pattern 1: Generalized Star (2)

Extending along "Port-Number"

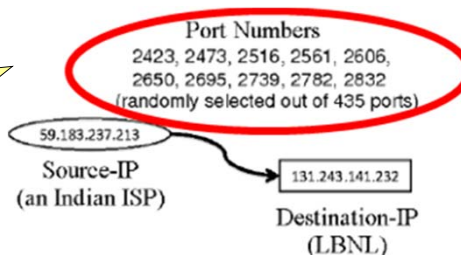


31

Pattern 1: Generalized Star (2)

Port numbers
used in packets
from the same
IP-src to the
same IP-dst

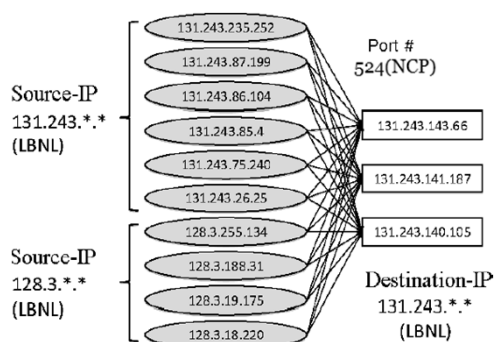
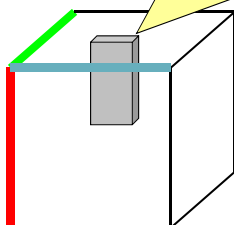
Port scanning
or P2P



32

Pattern 2: Generalized Bipartite-Core (1)

A 'plane' in a carefully
reordered tensor

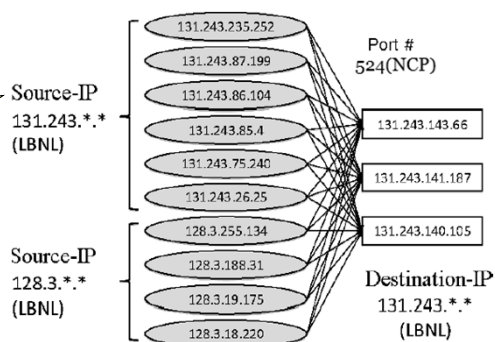


33

Pattern 2: Generalized Bipartite-Core (1)

IP-src's sending packets to the same IP-dst's & the same port

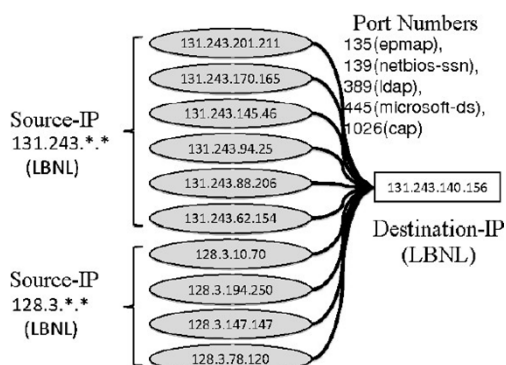
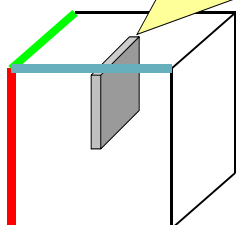
Clients talking to a shared server pool



34

Pattern 2: Generalized Bipartite-Core (2)

A 'plane' in a carefully reordered tensor

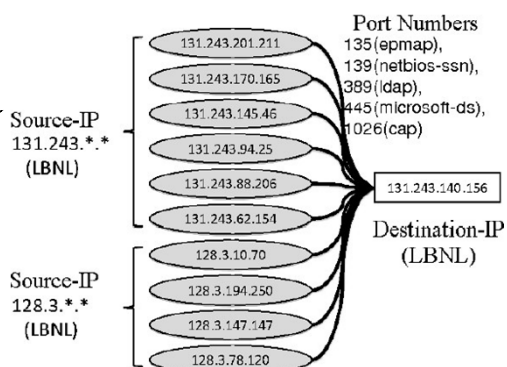


35

Pattern 2: Generalized Bipartite-Core (2)

IP-src's sending
packets over
multiple ports
to one IP-dst

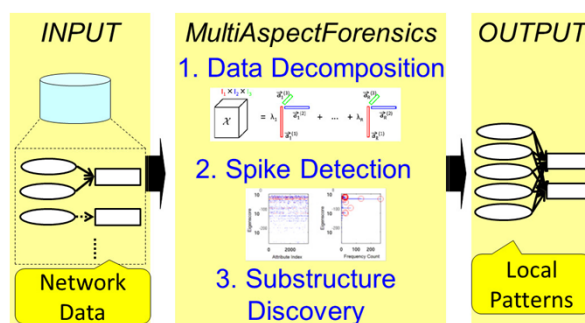
A multi-
purpose
windows server



36


M1: MultiAspectForensics

- Automatically detects novel patterns in heterogenous networks



37

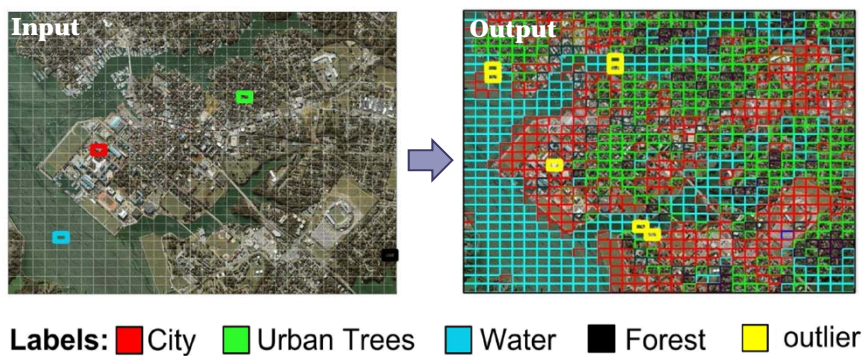
Thesis Outline

Mining	M1: MultiAspectForensics 
	M2: QMAS
Querying	Q1: Click Models
	Q2: C-DEM
	Q3: BEFH

38

QMAS: Mining Satellite Imagery (1)

- Low-labor labeling



39

QMAS: Mining Satellite Imagery (2)

- Low-labor labeling
- Identification of Representatives



40

QMAS: Mining Satellite Imagery (2)

- Low-labor labeling
- Identification of Representatives and Outliers



41

QMAS: Mining Satellite Imagery (2)

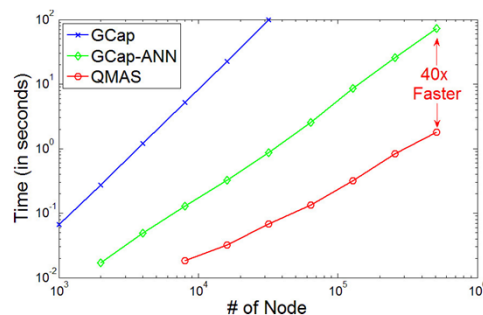
- Low-labor labeling
- Identification of Representatives and Outliers



42

QMAS: Mining Satellite Imagery (3)

- Low-labor labeling
- Identification of Representatives and Outliers
- Linear in time & space



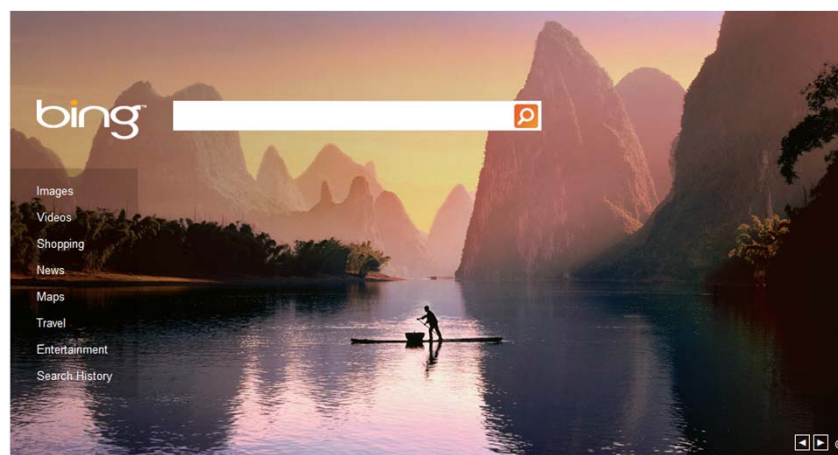
43

Thesis Outline

Mining	M1: MultiAspectForensics	✓
	M2: QMAS	✓
Querying	Q1: Click Models	
	Q2: C-DEM	
	Q3: BEFH	

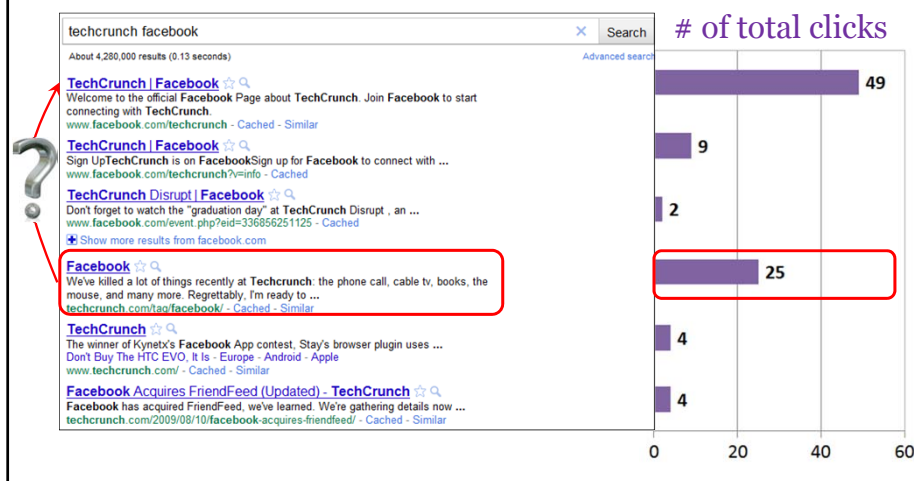
44

Web Search



45

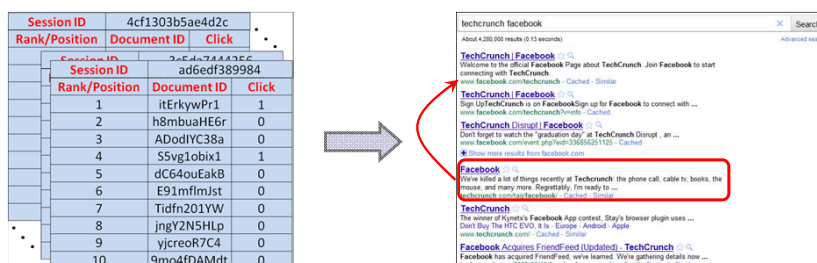
User Clicks as Quality Feedback



46

Motivation

- Leverage the signal from click data to improve search ranking.



47

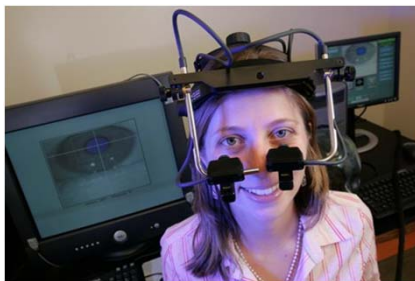
Click Through Rate (CTR)

- $\text{CTR} = \# \text{ of Clicks} / \# \text{ of Impressions}$

$$\text{Click} / (\text{Click} + \text{No Click})$$

48

Position Bias



49

Relevance of Web Document

- **Relevance = CTR @ Position 1**


$$= \frac{\# \text{ Clicks @ Position 1}}{\# \text{ Impressions @ Position 1}}$$

50

Problem Definition

- Estimate the relevance of web documents given clicks and their positions.

Session ID	4cf1303b5ae4d2c		
Rank/Position	Document ID	Click	
			...
Session ID	3e5da74442f6		
Session ID	ad6edf389984		
Rank/Position	Document ID	Click	
1	itErkywPr1	1	
2	h8mbuaHE6r	0	
3	ADodlYC38a	0	
4	S5vg1obix1	1	
5	dC64ouEakB	0	
6	E91mflmJst	0	
7	Tidfn201YW	0	
8	jngY2N5Hlp	0	
9	yjcreoR7C4	0	
10	9mo4fDAMdt	0	
			...



Document ID	Relevance
9mo4fDAMdt	0.12
ADodlYC38a	0.63
E91mflmJst	0.03
S5vg1obix1	0.36
Tidfn201YW	0.64
Zfds832lkkc	0.36
dC64ouEakB	0.88
h8mbuaHE6r	0.21
itErkywPr1	0.09
⋮	⋮

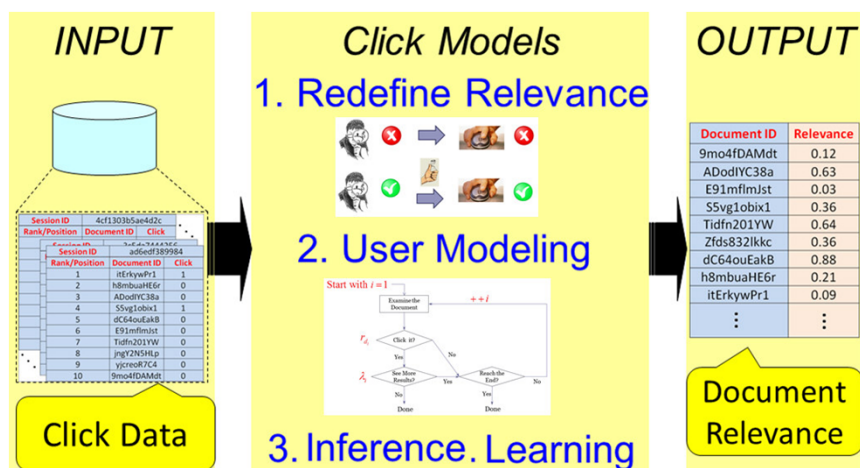
51

Design Goals / Constraints

- **Scalable**: single-pass, easy to parallel.
- **Incremental**: real-time updates possible.
- **Accurate**: consistent with past and future observations.

52

Approach



53

User Behavior Model

Facebook ☆ 🔍

We've killed a lot of things recently at Techcrunch: the phone call, cable tv, books, the mouse, and many more. Regrettably, I'm ready to ...
techcrunch.com/tag/facebook/ - Cached - Similar



Facebook Acquires FriendFeed (Updated) - TechCrunch ☆ 🔍

Facebook has acquired FriendFeed, we've learned. We're gathering details now ...
techcrunch.com/2009/08/10/facebook-acquires-friendfeed/ - Cached - Similar



Facebook | CrunchBase Profile 🔍

(techcrunch.com) []; Elevation Invests Another \$120 Million in Facebook as that IPO Looks More Distant (techcrunch.com) []; Facebook adds comments to the ...
www.crunchbase.com/Companies - Cached - Similar



Facebook: TechCrunch Confirms \$35B Valuation - Tech Trader Daily ... 🔍

Nov 22, 2010 ... I don't know if everyone already saw this on Friday, but it's worth posting as a reminder for those who didn't: Michael Arrington with ...
blogs.barrons.com/.../facebook-techcrunch-confirms-35b-valuation/ - Cached



54

Last Clicked Position

techcrunch facebook

×

Search

About 4,280,000 results (0.13 seconds) [Advanced search](#)

TechCrunch | Facebook ☆ 🔍
Welcome to the official Facebook Page about TechCrunch. Join Facebook connecting with TechCrunch.
www.facebook.com/techcrunch - Cached - Similar

TechCrunch | Facebook ☆ 🔍
Sign Up TechCrunch is on Facebook Sign up for Facebook to connect with ...
www.facebook.com/techcrunch/?v=info - Cached

TechCrunch Disrupt | Facebook ☆ 🔍
Don't forget to watch the "graduation day" at TechCrunch Disrupt , an ...
www.facebook.com/event.php?eid=336856251125 - Cached

[Show more results from facebook.com](#)

Facebook ☆ 🔍
We've killed a lot of things recently at Techcrunch: the phone call, cable tv, mouse, and many more. Regrettably, I'm ready to ...
techcrunch.com/tag/facebook/ - Cached - Similar

TechCrunch ☆ 🔍
The winner of Kynet's Facebook App contest, Stay's browser plugin uses ...
Don't Buy The HTC EVO, It Is - Europe - Android - Apple
www.techcrunch.com/ - Cached - Similar

Facebook Acquires FriendFeed (Updated) - TechCrunch ☆ 🔍
Facebook has acquired FriendFeed, we've learned. We're gathering details now ...
techcrunch.com/2009/08/10/facebook-acquires-friendfeed/ - Cached - Similar

55

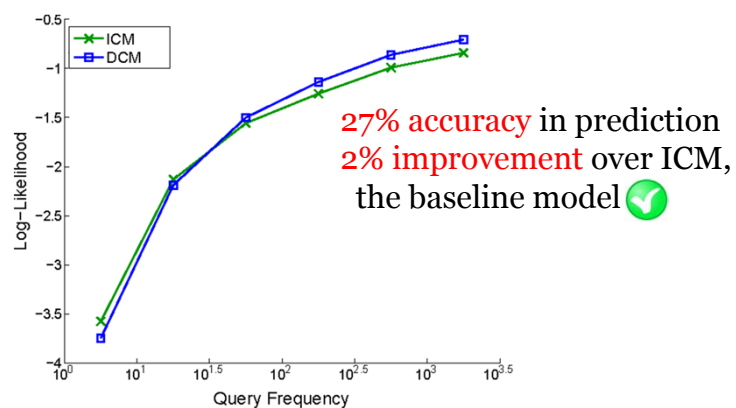
Empirical Results

- Click data after pre-processing
 - 110K distinct queries, 8.8M query sessions.
- Training time: <6 mins ✓
- Online update: ✓
 - Bump impression and click counters
 - No data retention required

56

Empirical Results

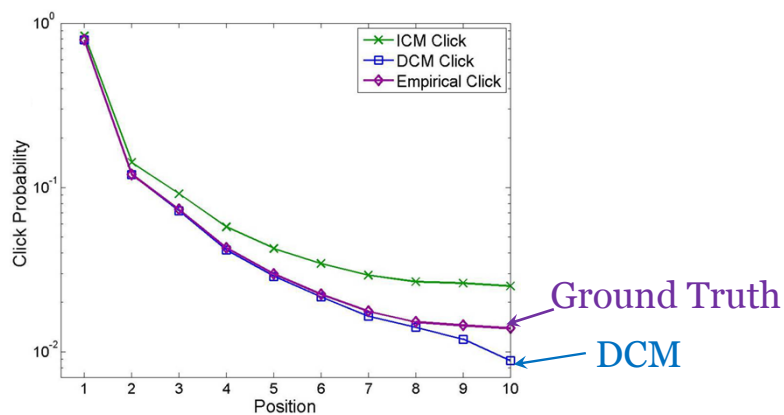
- Higher log-likelihood indicates better quality.



57

Empirical Results

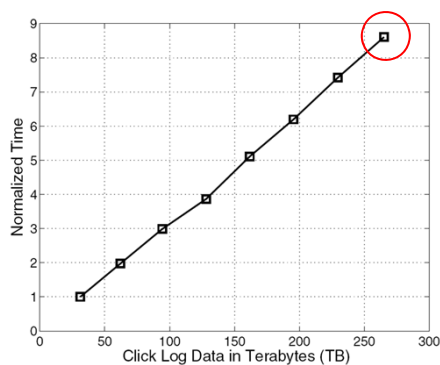
- Position-bias visualized



58

Scaling to Terabytes

- 265TB** data, 1.15B document relevance results, running time on wall clock ~ **3 hours**



59

Q1: Click Models

- A statistical approach to leveraging click data for better ranking aware of position-bias.
- They are **incremental**, more **accurate** than the baseline, scaling to almost **petabyte-scale** data.

60

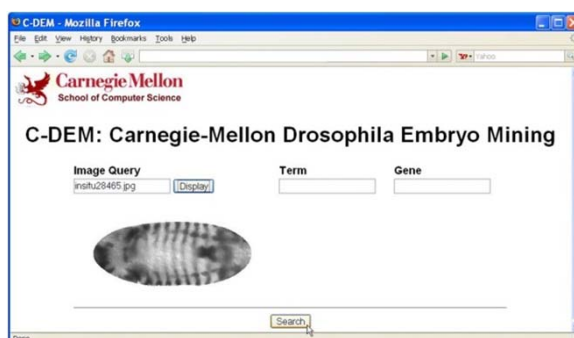
Thesis Outline

Mining	M1: MultiAspectForensics	✓
	M2: QMAS	✓
Querying	Q1: Click Models	✓
	Q2: C-DEM	
	Q3: BEFH	

61

Q2: C-DEM

- A flexible query interface for 3-mode data: images, genes, annotation terms.



62

Q2: C-DEM

Images →

Terms →

Genes →

Relevant Image	Score	Relevant Image	Score
	0.141954		0.105069
	0.092182		0.014665
	0.011089		0.010085

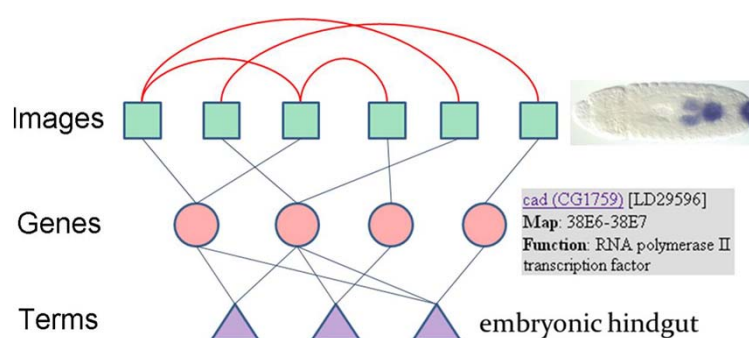
Relevant Term	Score	Relevant Gene Score
embryonic dorsal epidermis	0.006391	0.072133
embryonic ventral epidermis	0.006176	Lac
embryonic hindgut	0.005796	Cu-P60A
embryonic foregut	0.005372	CG17218
embryonic larval tracheal system	0.005169	CG5532
embryonic head epidermis	0.004742	btp

search by checking the box or Return to the query page!

63

Q2: C-DEM

- Solution: random walk with restart on graphs.



64

Thesis Outline

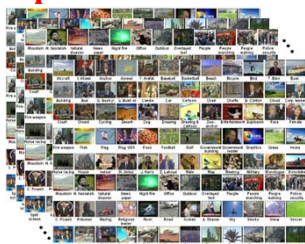
Mining	M1: MultiAspectForensics	✓
	M2: QMAS	✓
Querying	Q1: Click Models	✓
	Q2: C-DEM	✓
	Q3: BEFH	

65

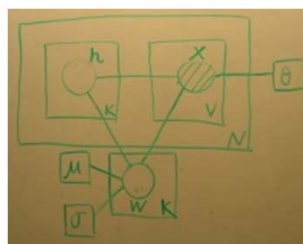
Q3: BEFH (1)

- Bayesian exponential family harmonium
- Deriving topical representations for multimedia corpora (e.g., *video snapshots* and *captions*)

Input



Model

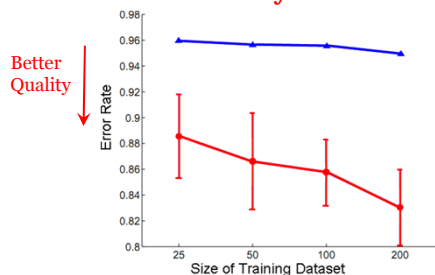


66

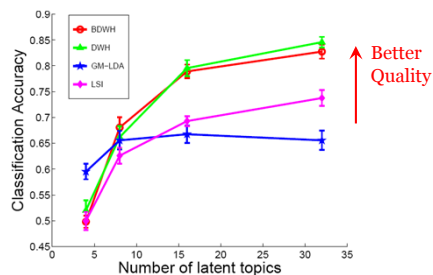
Q3: BEFH (2)

- Bayesian exponential family harmonium
- Deriving topical representations for multimedia corpora (e.g., *video snapshots* and *captions*)

Validation – Synthetic Data



Validation – TRECVID Data



67

Thesis Outline

Mining	M1: MultiAspectForensics	✓
	M2: QMAS	✓
Querying	Q1: Click Models	✓
	Q2: C-DEM	✓
	Q3: BEFH	✓

68

Conclusion

- Data-driven research under the theme of **pattern mining** and **similarity querying**.

73

Thank You!

- <http://www.cs.cmu.edu/~fanguo/dissertation/>




74

RTW Knowledge Base Query Interface

This is a simple interface to query the knowledge base from [Read The Web \(RTW\) Project](#), which consists millions of (entity, relation, value) triplets, also known as facts, such as ("pittsburgh", "citylocatedinstate", "pennsylvania"). It supports the following two functions:

1. Fact Browsing: list all facts for a given query word and its role;
2. Similarity Query: based on a fast approximation of personalized pagerank.

Query word: ☐ Show query suggestion

Choose a task for the query word: 

Fact Browsing
 Similarity Query

Query Results: