Scalable Processing of Spatial-Keyword Queries

Walid G. Aref
Purdue University and
Alexandria University-Egypt

Ahmed Mahmood
Google LLC.

PURDUE
UNIVERSITY

Outline

- Introduction and Background
- Querying Spatial-Keyword Data
- Scalable Spatial-Keyword Systems
- Research Directions

Introduction and Background

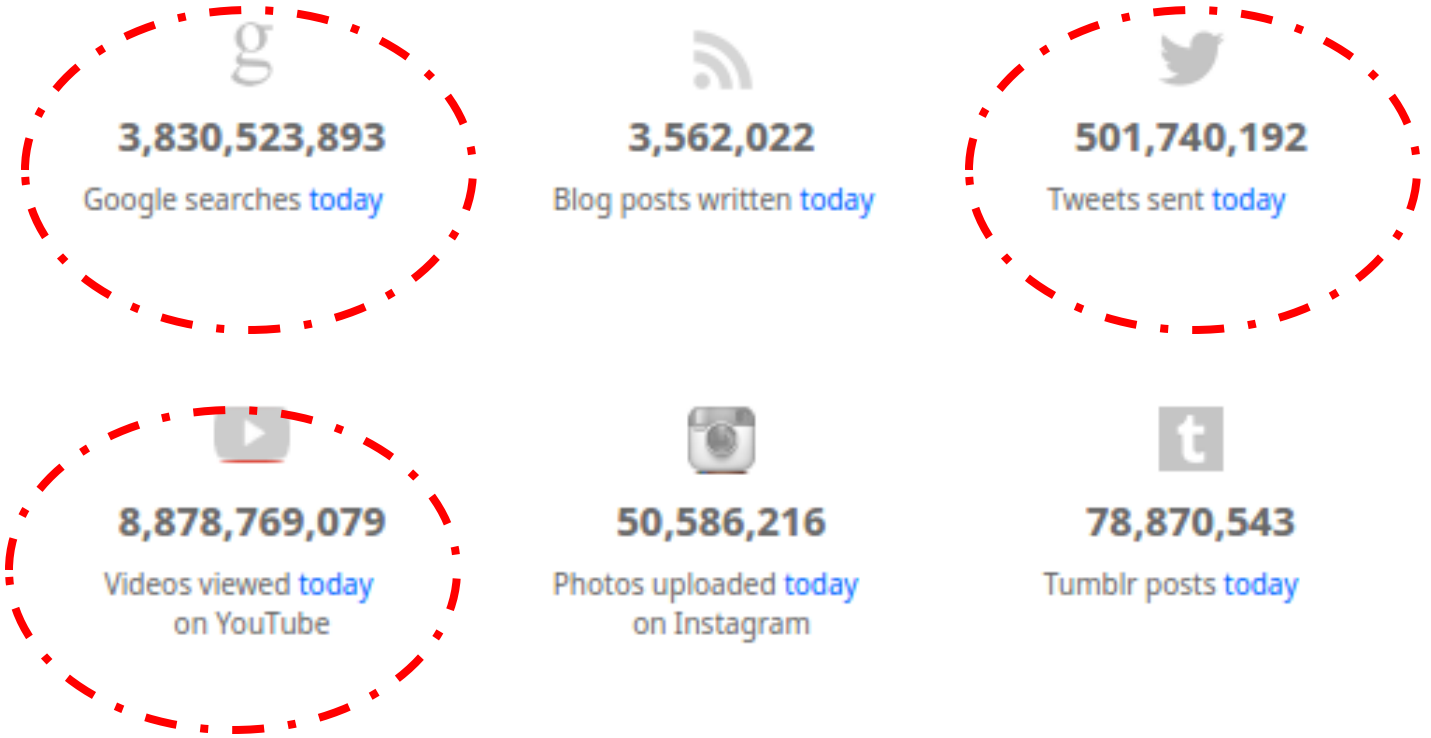
- The Scale of Spatial-Keyword Data
- Spatial-Keyword Applications

Big Spatial-Keyword Data

- Spatial-keyword data is composed of:
  - Geo-location
  - Textual content, i.e., set of keywords
  - Other attributes, e.g., timestamp, user identifier
- Large amounts of spatial-keyword data are being generated
  - Location-aware devices, e.g., smart-phones
  - Social networks
  - Micro-blogging applications

Spatial-Keyword Data is Everywhere

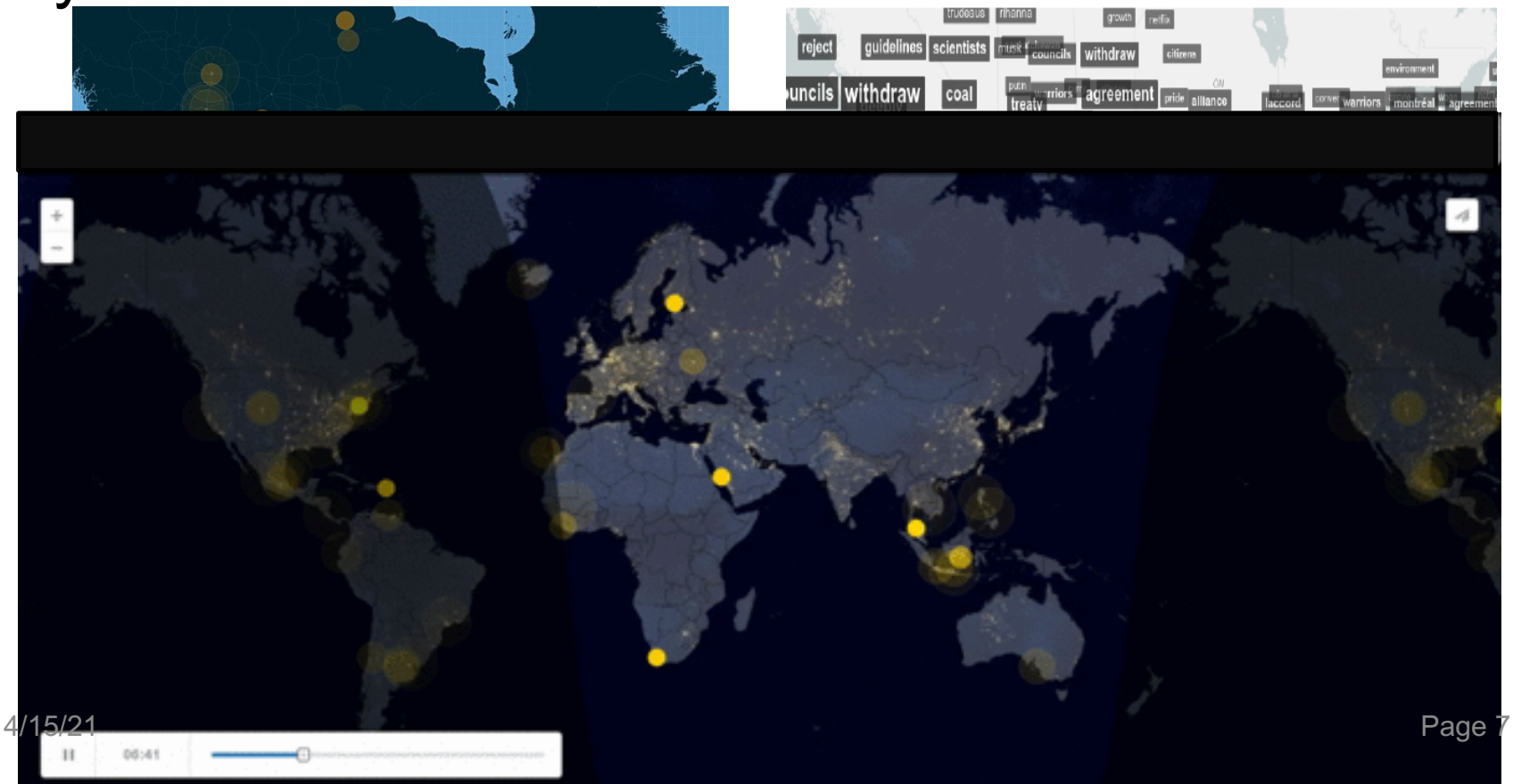
The Scale of Spatial-Keyword Data



http://www.internetlivestats.com/

Observations on Spatial-Keyword Data

- The nonuniform and dynamic nature of spatial-keyword data



General Purpose Big Data Systems

- These systems are not equipped with spatial-keyword-specific indexing and query processing algorithms

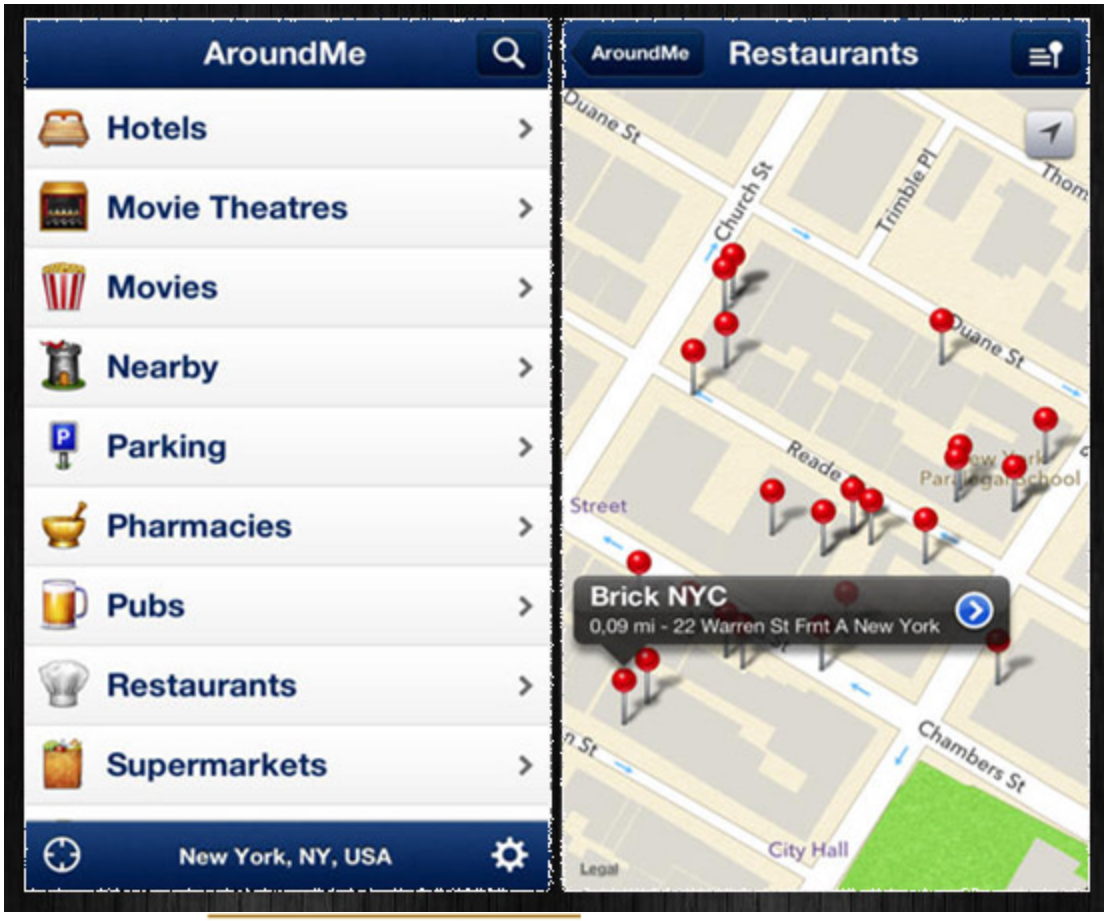
Introduction and Background

- The Scale of Spatial-Keyword Data
- Spatial-Keyword Applications

Ad Targeting

- Online location-based advertising campaigns
- Location-based advertising, e.g., mobile e-coupons
- Also known as location-aware publish/subscribe systems

Identification of Nearby Attractions

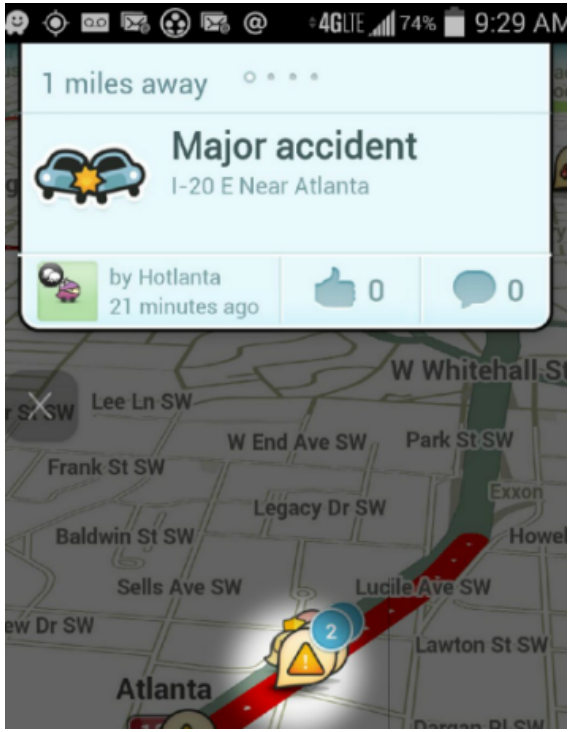


Trip Planning

- A tourist wants to identify ***groups*** of points that are close to each other

Traffic and Navigation Systems

- Waze



Micro-Blogs Analysis

https://www.trendsmap.com/

Outline

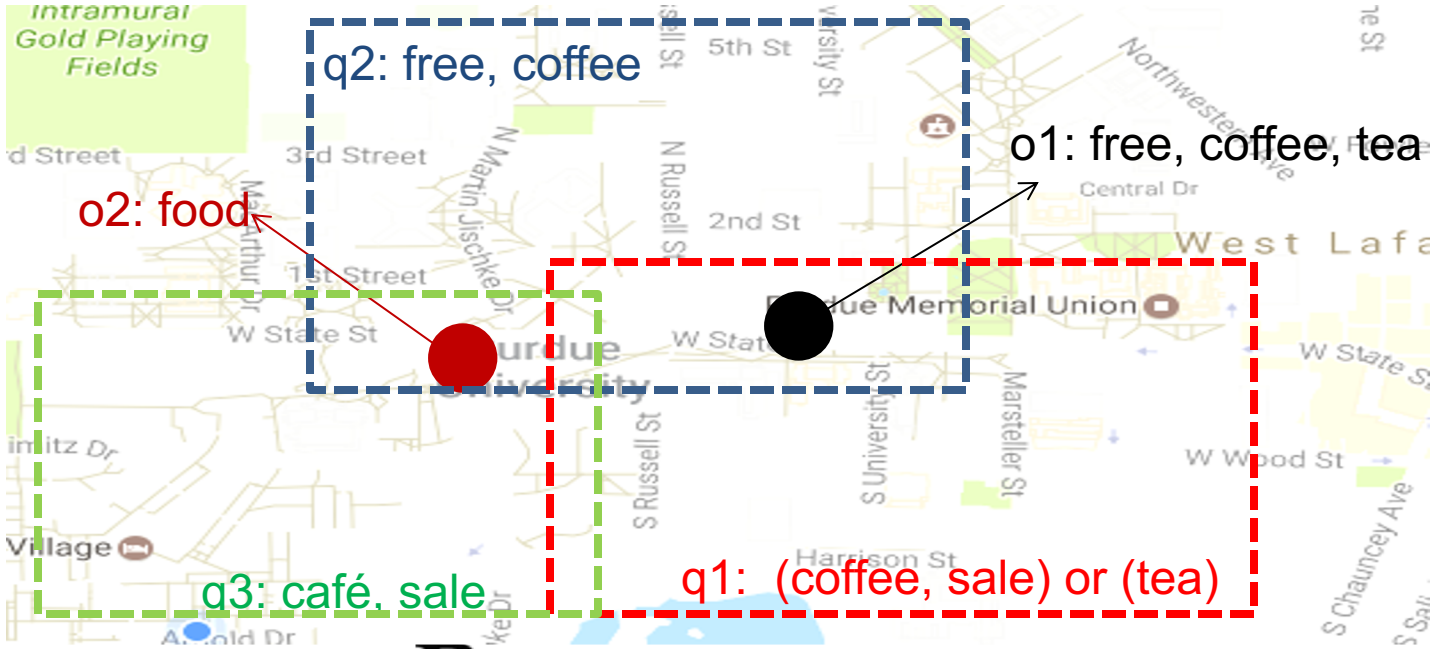
- Introduction and Background
- Querying Spatial-Keyword Data
- Scalable Spatial-Keyword Systems
- Research Directions

Querying Spatial-Keyword Data

- Spatial-Keyword Queries
- Spatial-Keyword Query Languages

The Spatial-Keyword Filter Query

- Identifies objects located inside the spatial range of the query and matches the textual criteria of the query



The Spatial-Keyword Top-K Query

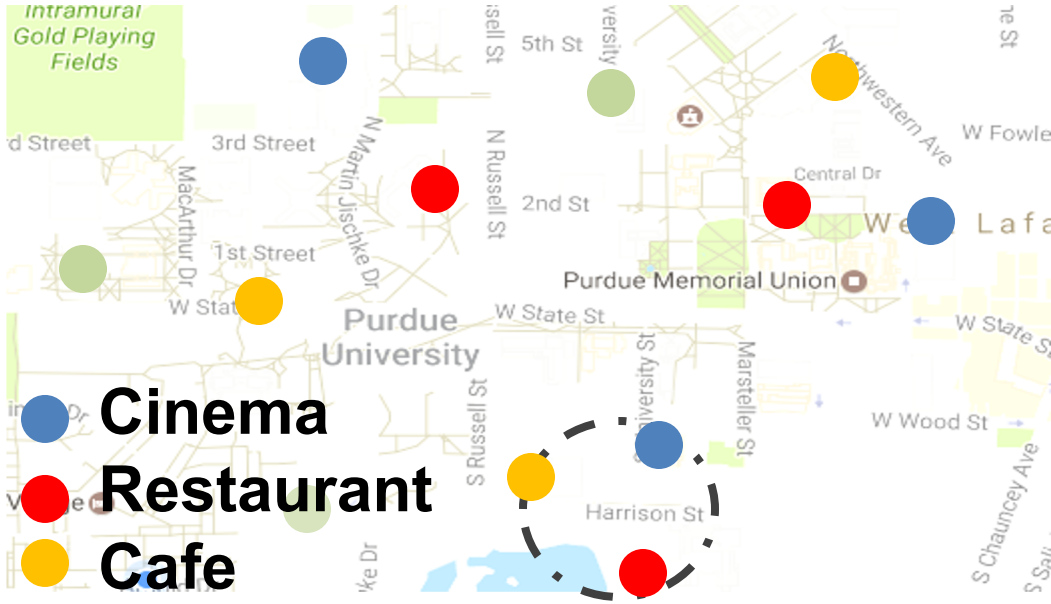
- Retrieve the top-k spatial-keyword objects ranked based on a function of:
  - Distance between the object and the query location
  - Similarity between the keywords of the object and the query
- Example: Find the top three tuples that have any of the keywords: restaurant, seafood

The Boolean-kNN Query

- A special version of the spatial-keyword top-K query
  - Identify all objects that contain the query keywords
  - Rank these objects based on the distance between objects and the location of the query

The m-Closest Keywords (mCK) Query

- Identify the group of spatial-keyword objects that collectively cover a set of keywords and have a minimum diameter

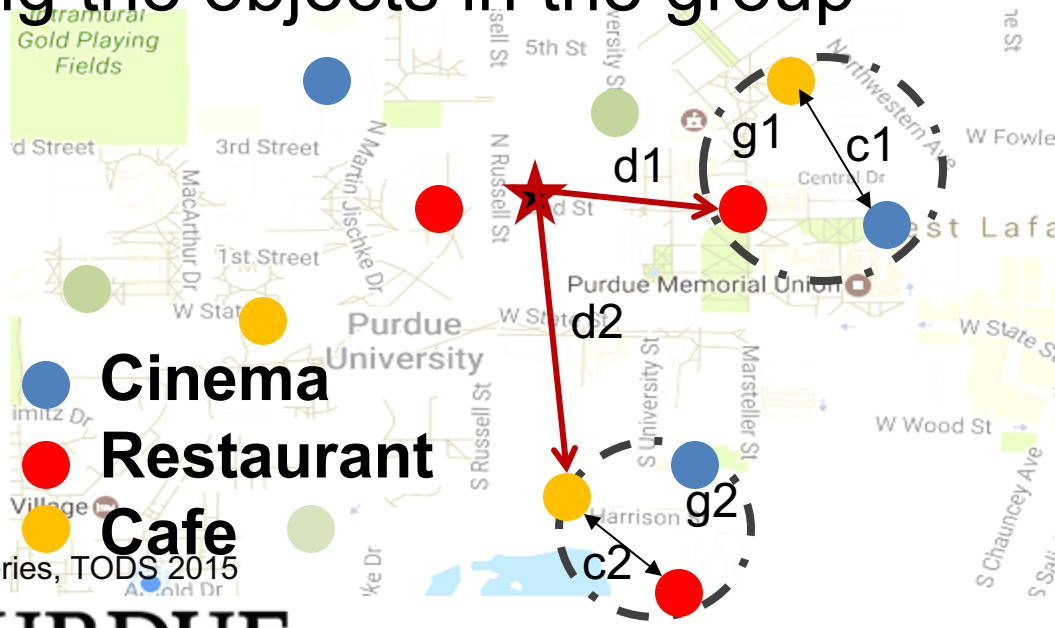


Zhang et al., Keyword search in spatial databases: Towards searching by document, ICDE 2009
Zhang et al., Locating mapped resources in web 2.0, ICDE 2010
Guo et al., Efficient algorithms for answering the m-closest keywords query, SIGMOD 2015

The Collective Spatial-Keyword Group Query (CoSKQ)

Find a group of objects that collectively cover the keywords of the query and minimize a function of:

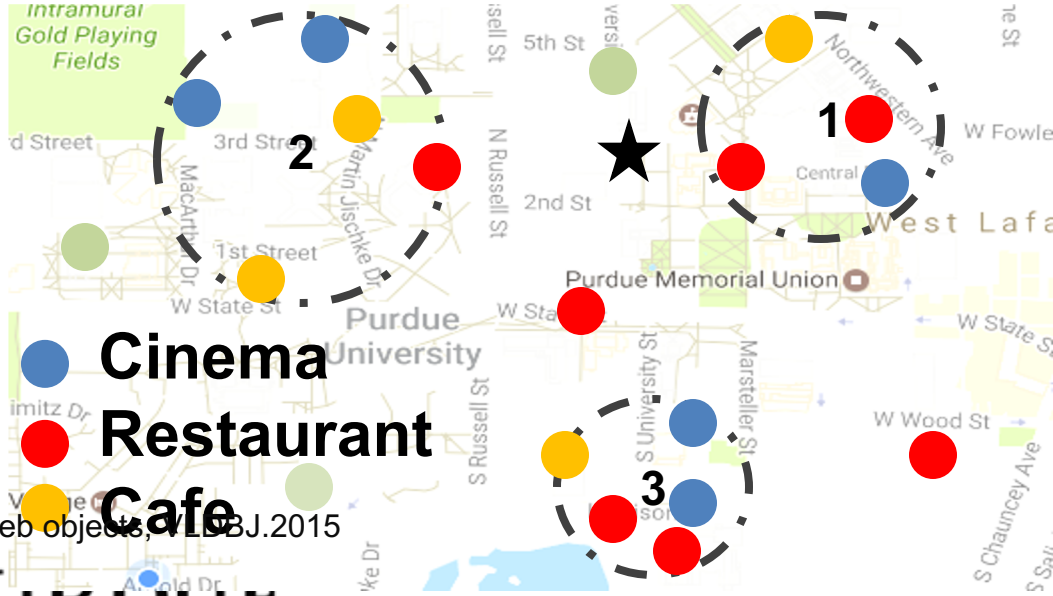
- The distance between the query and the group
- The distances among the objects in the group



Cao et al., Efficient processing of spatial group keyword queries, TODS 2015

The Top-k Groups Query

- Identifies multiple groups of spatial-keyword objects that collectively cover the keywords of the query
- Groups are ranked based on a function of
  - Distance of the group to the location of the query
  - The distances among the objects in the group
  - The multiplicity of objects covering the same keywords
- A group may contain multiple objects covering the same keyword



Skovsgaard et al., Finding top-k relevant groups of spatial web objects, VLDBJ.2015

Variations of Spatial-Keyword Queries

- Spatial-Keyword Similarity Join
- Spatial-Keyword Skyline Query
- Spatial-Keyword queries on road-networks
- Continuous queries over spatial-keyword data streams
- Direction-aware spatial-keyword filter query

Querying Spatial-Keyword Data

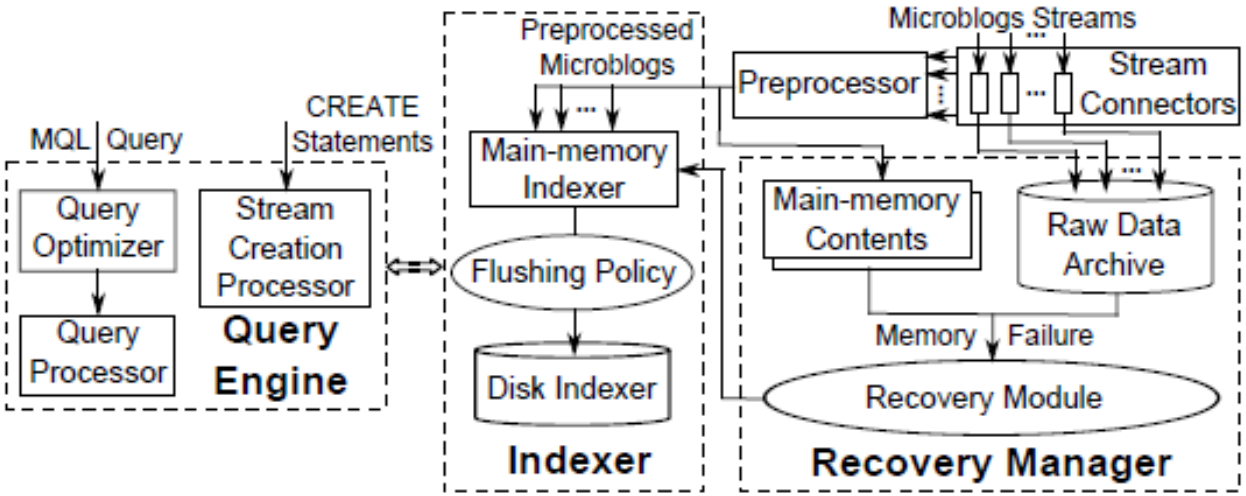
- Spatial-Keyword Queries
- Spatial-Keyword Query Languages

Querying Spatial-Keyword Data

- Spatial-Keyword Queries
- **Spatial-Keyword Query Languages**
  - **MQL**
  - Atlas

MQL Micro-Blogs Query Language

- A micro-blogs management system along with a micro-blogs query language



Magdy et.al., Towards a microblogs data management system, MDM 2015

MQL Micro-Blogs Query Language

```
* SELECT [CONTINUOUS] attr_list
FROM stream_name1 [,stream_name2,...]
[WHERE condition]
ORDER BY F(arg_list)
LIMIT k
ON {LAST T {MINUTES|DAYS} | (T_start,T_end)}
```

```
* SELECT [CONTINUOUS] grouping_attr_list,
COUNT(attr_list)
FROM stream_name1 [,stream_name2,...]
[WHERE condition]
GROUP BY grouping_attr_list
LIMIT k
ON {LAST T {MINUTES|DAYS} | (T_start,T_end)}
```

MQL Micro-Blogs Query Language

- Example: Retrieve the most frequent 10 keywords from tweets in Ukraine since February 18, 2014:

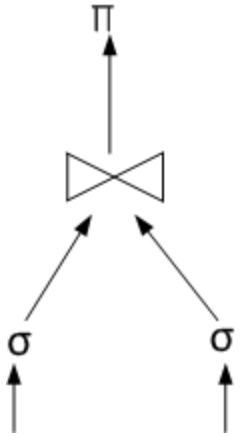
```
   SELECT CONTINUOUS keyword, COUNT(*)
FROM twitter_stream
WHERE location WITHIN (52,44.7,39.91,21.8)
GROUP BY keyword
LIMIT 10 ON ("18 Feb 2014",∞)
```

Querying Spatial-Keyword Data

- Spatial-Keyword Queries
- **Spatial-Keyword Query Languages**
  - MQL
  - **Atlas**

The Atlas Query Language Spatial-Keyword Building Blocks

- Design building-block operators that can be composed to realize complex spatial-keyword queries
  - Similar to the relational SELECT, PROJECT, JOIN building-block operators



Mahmood et al., Atlas: on the expression of spatial-keyword group queries using extended relational constructs, SIGSPATIAL, 2016

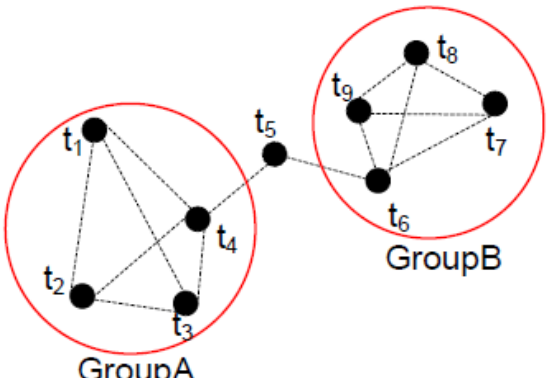
The Specification of Atlas

```
SELECT {*|attr1 [AS alias][,attr2,..]}
FROM source_name1 [,source_
name2, …]
[WHERE condition]
[ORDER BY F(arg_list)]
[LIMIT {k|condition}]
```

```
SELECT grp_attr1[AS alias][,grp_attr2,..}, AGGR_F [AS
alias](attr_list)
FROM source_name1 [,source_ name2, …]
[WHERE condition]
[PARTITION BY grp_att_list AS grp_alias]
[ORDER BY F(grp_arg_list)]
[LIMIT k]
[HAVING grp_condition]
```

Spatial-keyword Grouping Operators

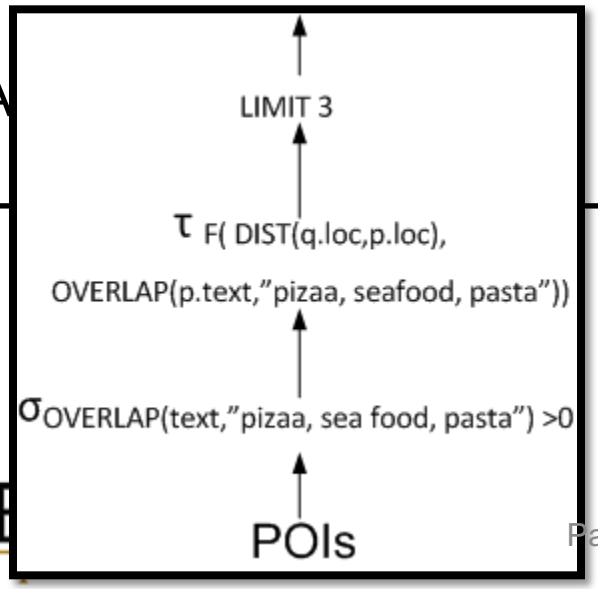
- CONDITIONAL LIMIT
  - Retrieves a list of items until a condition is satisfied
- PARTITION BY
  - Similar to the traditional *group-by*, yet returns the group tuples (not aggregates over the groups)
- WITHIN_DIST(D)
  - Identifies groups of tuples such that the distance between every pair of tuples within a group is upper-bounded by D



Example 1: Spatial-Keyword TOP-K Query

- Retrieve the top-k spatial-keyword objects ranked by a function of:
  - Distance between the object and the query location
  - Similarity between the keywords of the object and the query
- Example: Find the top three tuples that have any of the keywords: pizza, seafood, pasta
- Q.keywords and Q.loc

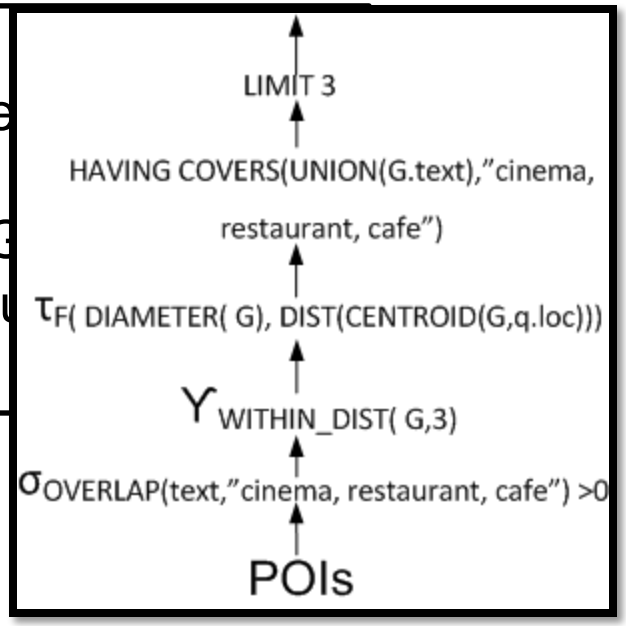
```
SELECT * FROM source S, query Q
WHERE OVERLAP(S.txt, Q.keywords)>0
ORDER BY  F(DIST(S.loc, Q.loc), OVERLA
LIMIT 3
```



Example 2: Spatial-Keyword Group Query

Find groups of objects that collectively contain the keywords “cinema, restaurant, cafe” such that objects in a group are within 3 miles of each other. The groups are to be **ranked by a function of the groups’ inter-object distances and the distance from each of the groups to a specific location**

```
SELECT * FROM source S
WHERE OVERLAP(S.txt, “cinema, restaurant, cafe
PARTITION BY  WITHIN_DIST(S.loc,3) AS G
ORDER BY  F(DIAMETER(G), DIST(CENTROID(G
HAVING CONTAINS(UNION(G.txt), “cinema, restau
LIMIT  3
```



Outline

- Introduction and Background
- Querying Spatial-Keyword Data
- Scalable Spatial-Keyword Systems
- Research Directions

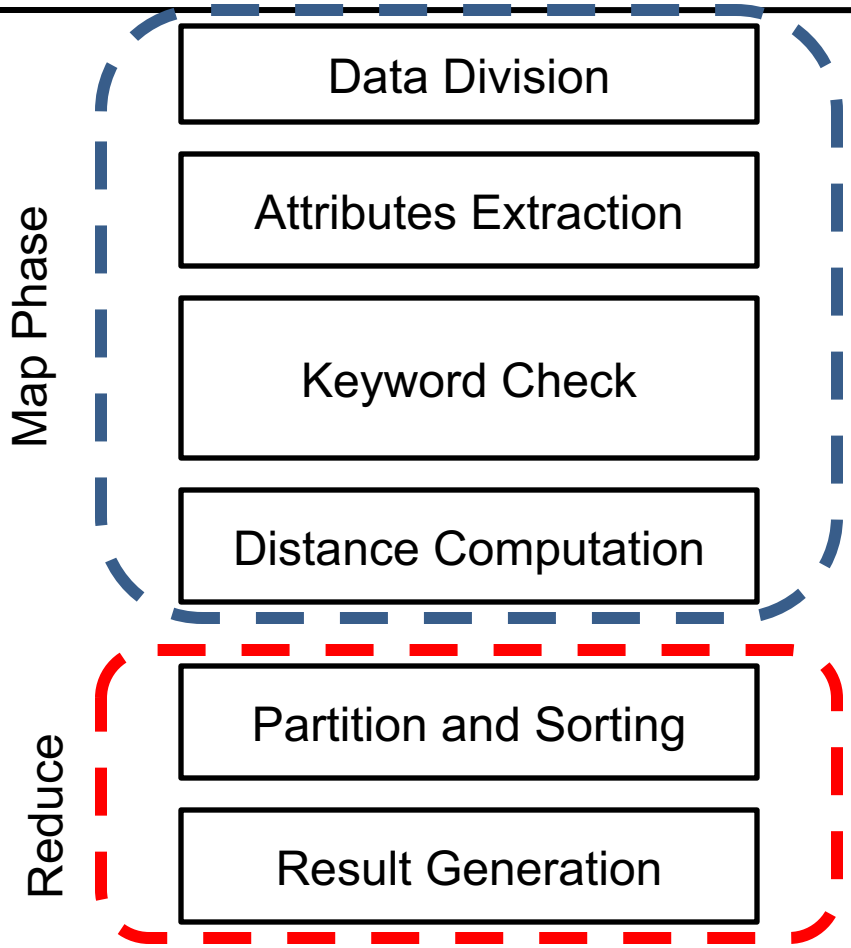
Outline

- Introduction and Background
- Querying Spatial-Keyword Data
- **Scalable Spatial-Keyword Systems**
- Research Directions

Scalable Spatial-Keyword Systems

- Query-specific algorithms
- Extensions to general-purpose big data systems
- Spatial-keyword-only systems

Boolean kNN Processing over MapReduce



- A map-reduce algorithm to answer the Boolean kNN spatial-keyword query over unstructured data

Li et al., Evaluating spatial keyword queries under the mapreduce framework, DASFAA 2012

Big Spatial-Keyword Processing Systems

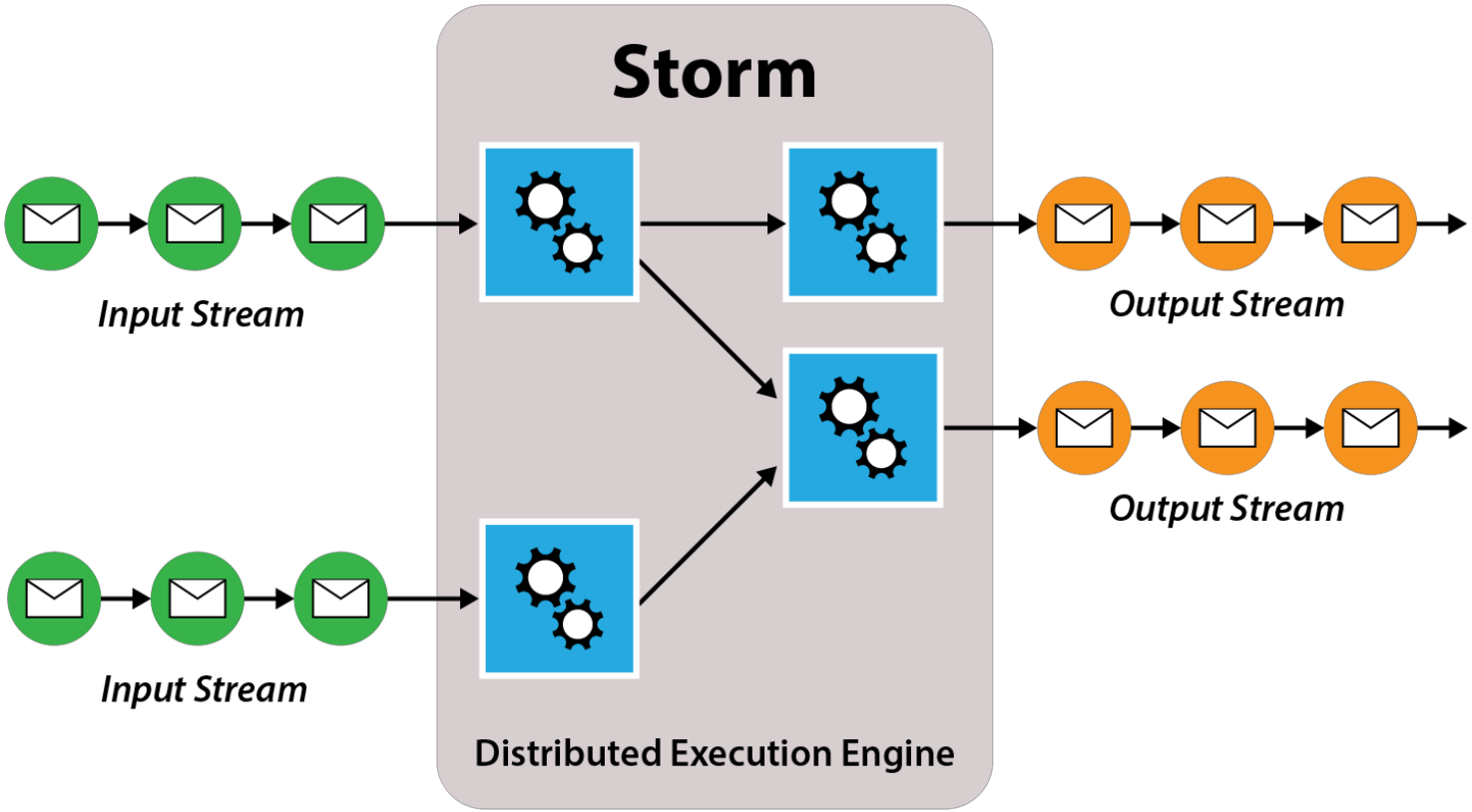
- Query-Specific Algorithms
- Extensions to general-purpose big data systems
- Spatial-keyword-only systems

Tornado

- Adaptive and Distributed System for Spatial-Keyword Stream Processing
  - Extend the general-purpose Storm streaming system
  - Distribute the input streams to multiple worker processes

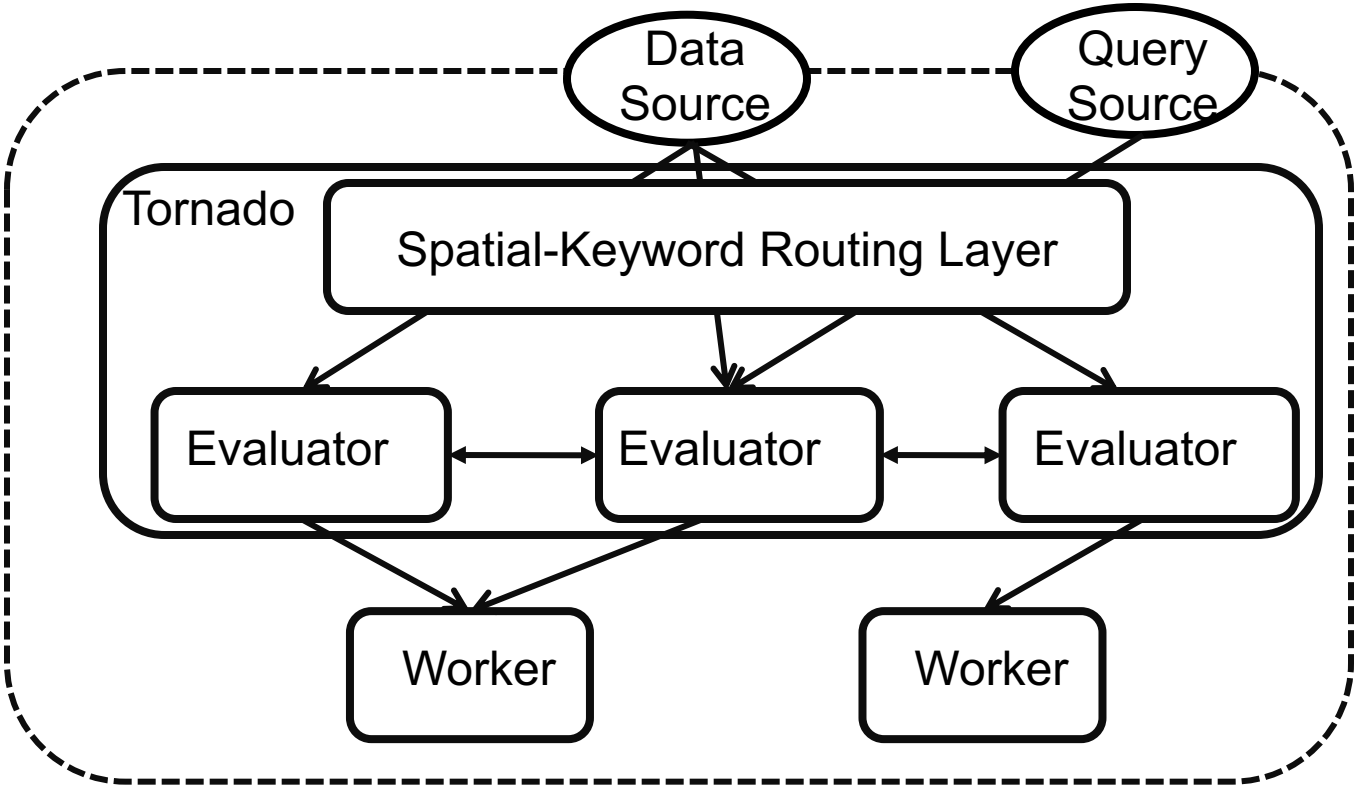
Mahmood et al., Adaptive processing of spatial-keyword data over a distributed streaming cluster, SIGSPATIAL 2018

The Storm Streaming System



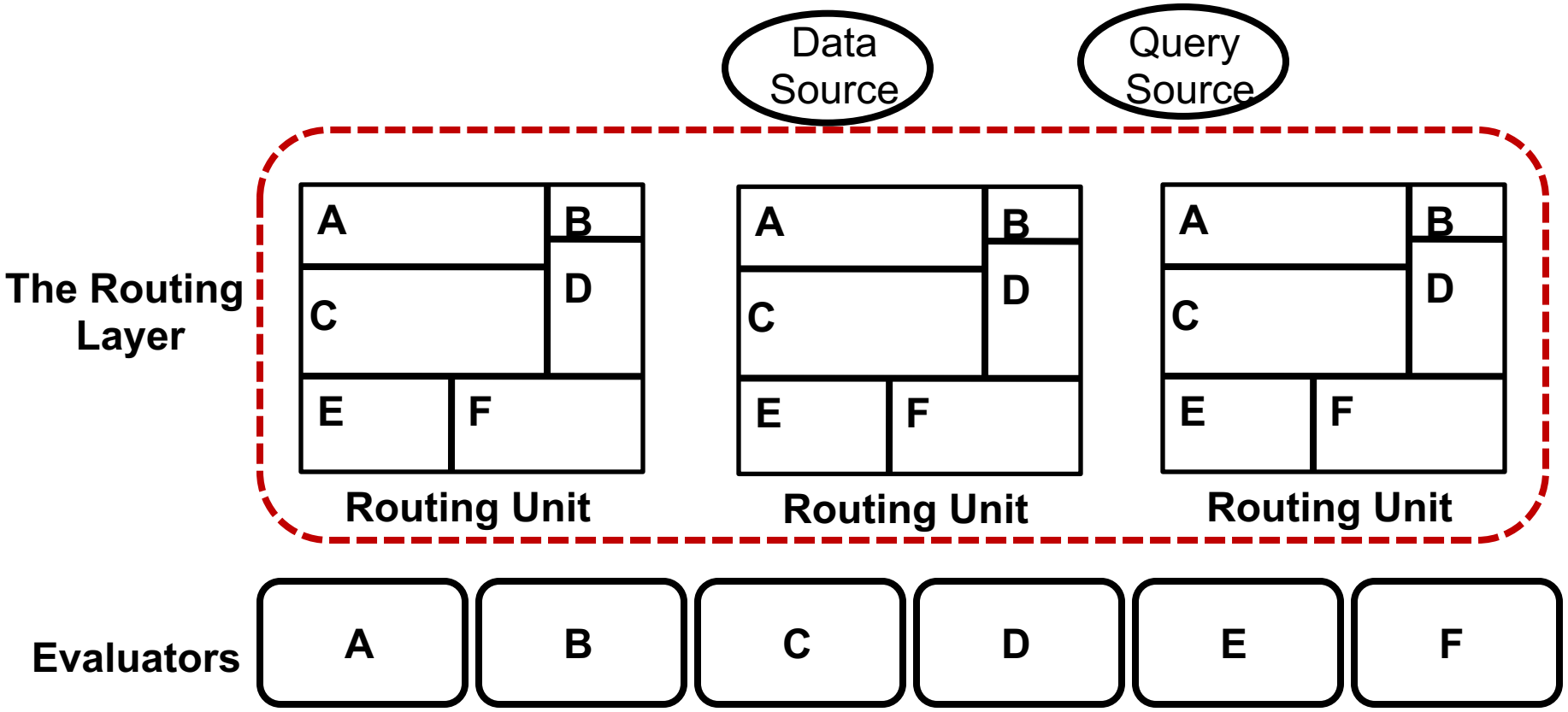
http://www.business-software.com/wp-content/uploads/2014/09/Storm.png

Tornado: Adaptive and Distributed System for Spatial-Keyword Stream Processing



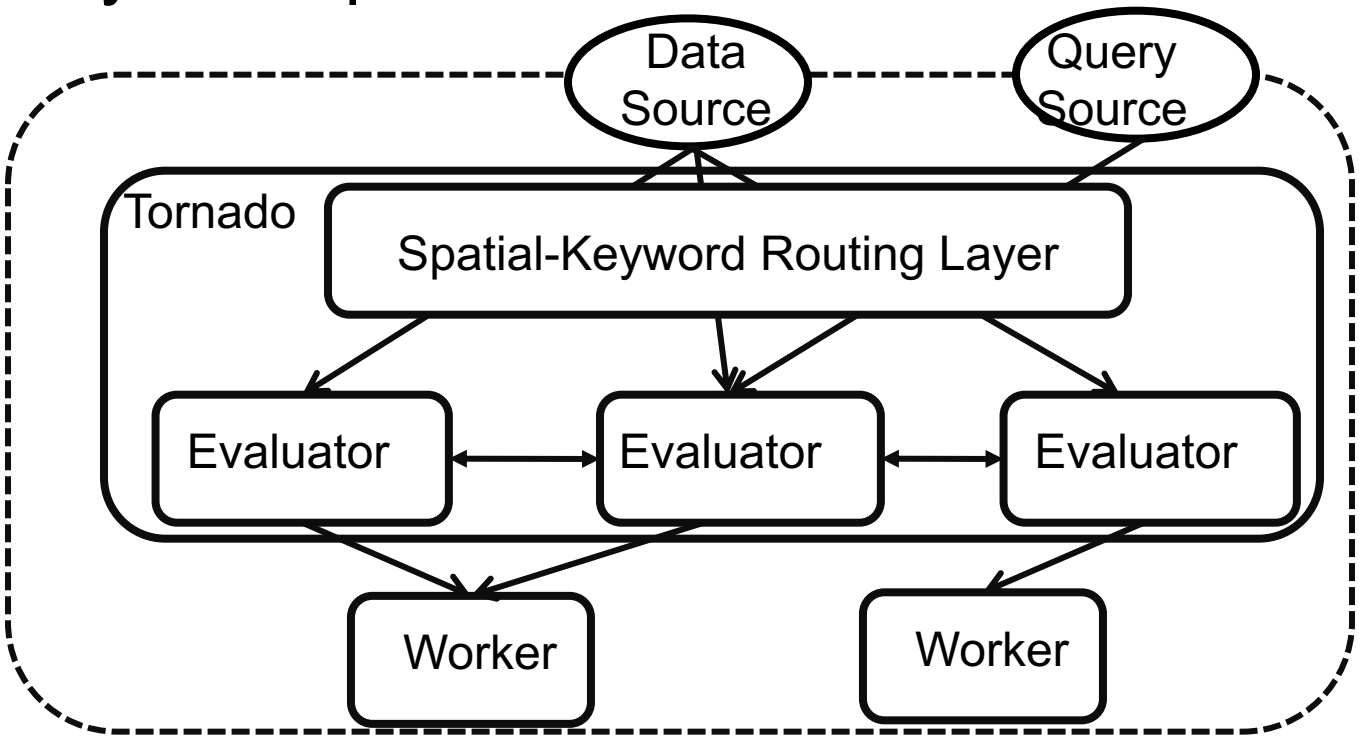
The Routing Layer

- Partition the space into non-overlapping rectangles
- Identify evaluators to which data objects and queries belong to



Tornado: Adaptive and Distributed System for Spatial-Keyword Stream Processing

- Evaluators use FAST* to internally index continuous spatial-keyword queries

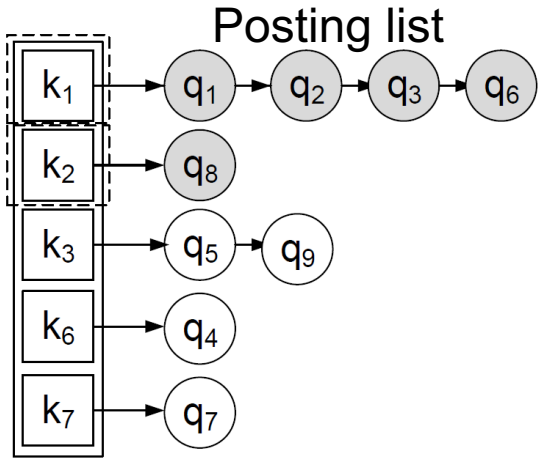


Limitations of Existing Spatial-Keyword Indexes

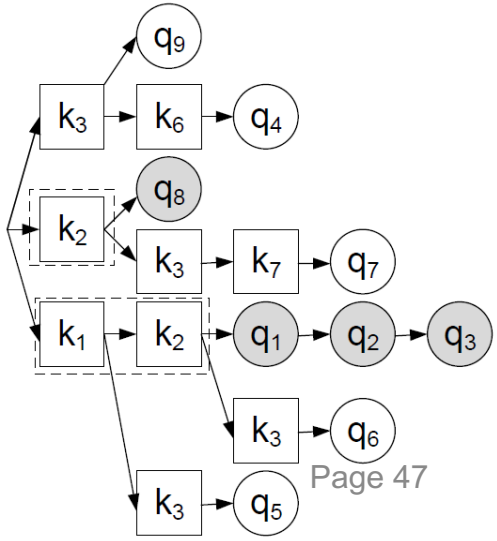
- Integrate a spatial index with a textual index
  - Do not account for the popularity of keywords across different regions
- Spatial Indexes
  - R-tree
    - High update overhead
  - Grid
    - High memory requirements due to replication of textual structures in multiple grid cells

Limitations of Existing Spatial-Keyword Indexes

- Textual Indexes
  - Ranked inverted list
    - Poor search performance for long posting lists



  - Ordered keyword trie
    - Similar to the traditional trie
    - High memory requirements
    - Requires having a total order on the keywords
    - Requires knowing the entire vocabulary

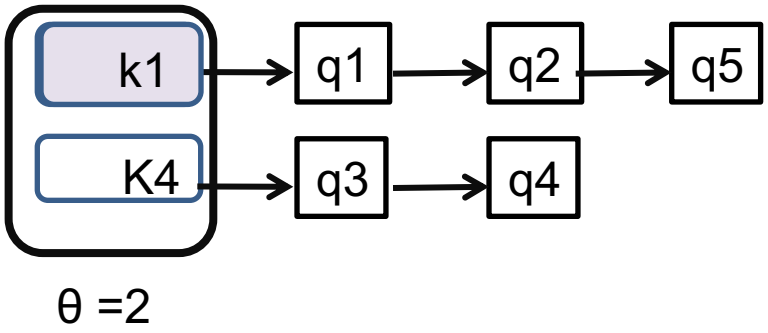


The Adaptive Keyword Index (AKI)

- A textual index that accounts for the popularity of keywords
- Starts as a ranked inverted list (RIL)
- The frequent-keyword threshold θ is used to prevent having long posting lists
- A query is attached to its least frequent keyword
- Frequencies of keywords are not known apriori
  - progressively maintained

The Adaptive Keyword Index (AKI)

- When the number of queries attached to a keyword exceeds the frequent-keyword threshold θ, mark it as frequent

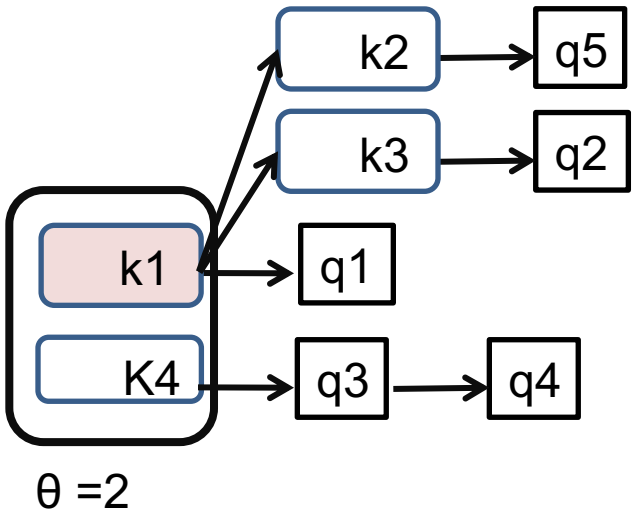


Frequencies Map

K1	3
...	
K4	2

The Adaptive Keyword Index (AKI)

- Use more keywords of queries to improve textual discrimination

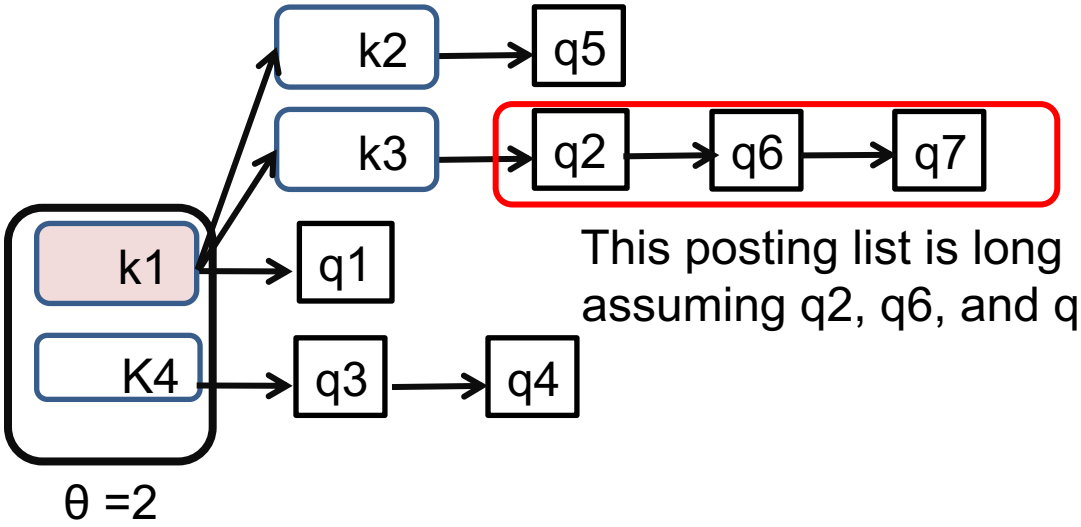


Frequencies Map

K1	3
...	
K4	2

The Adaptive Keyword Index (AKI)

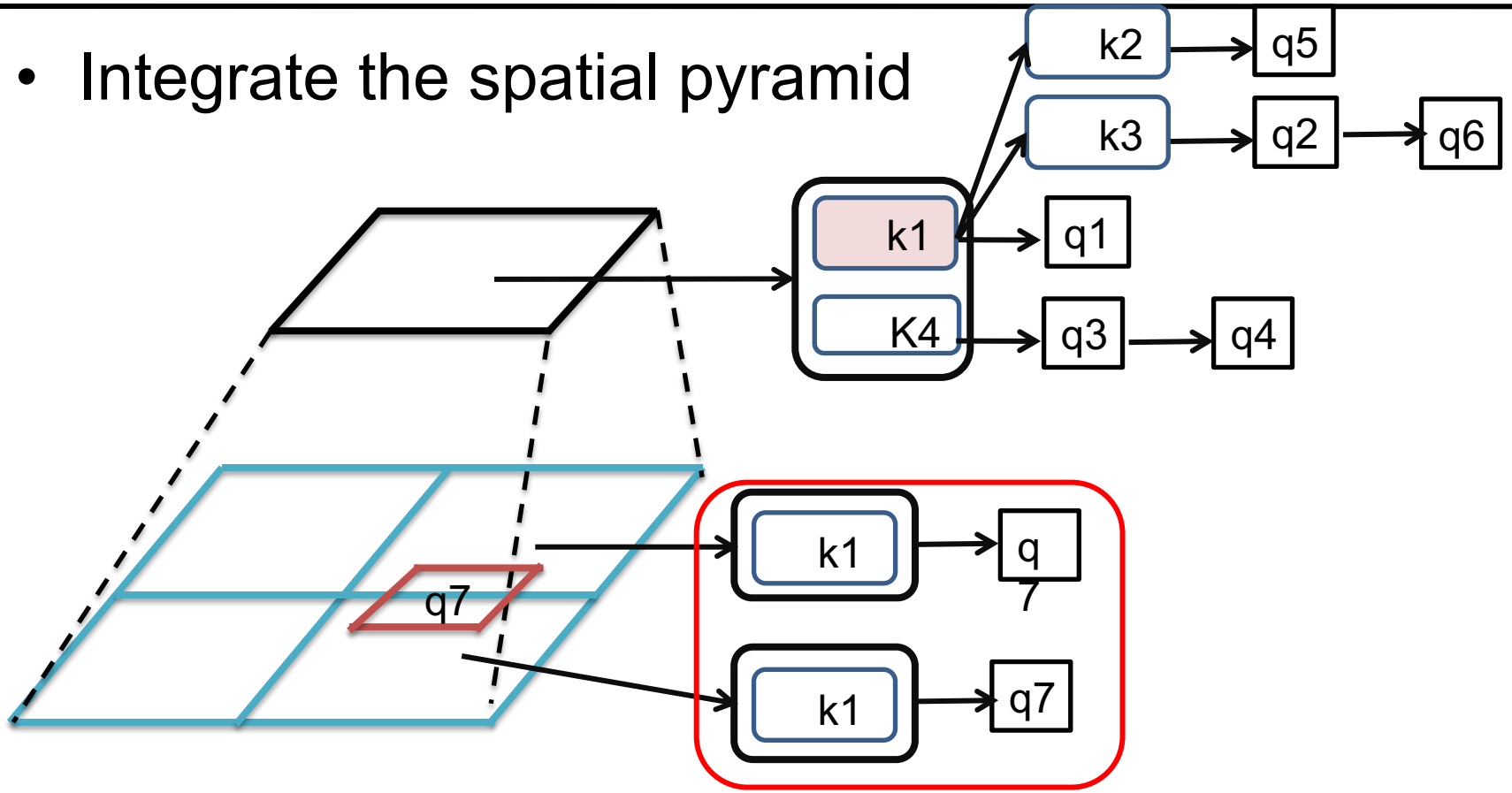
- What happens when many queries have the exact same keywords?



This posting list is long
assuming q2, q6, and q7 all have the keywords k1, k3

FAST

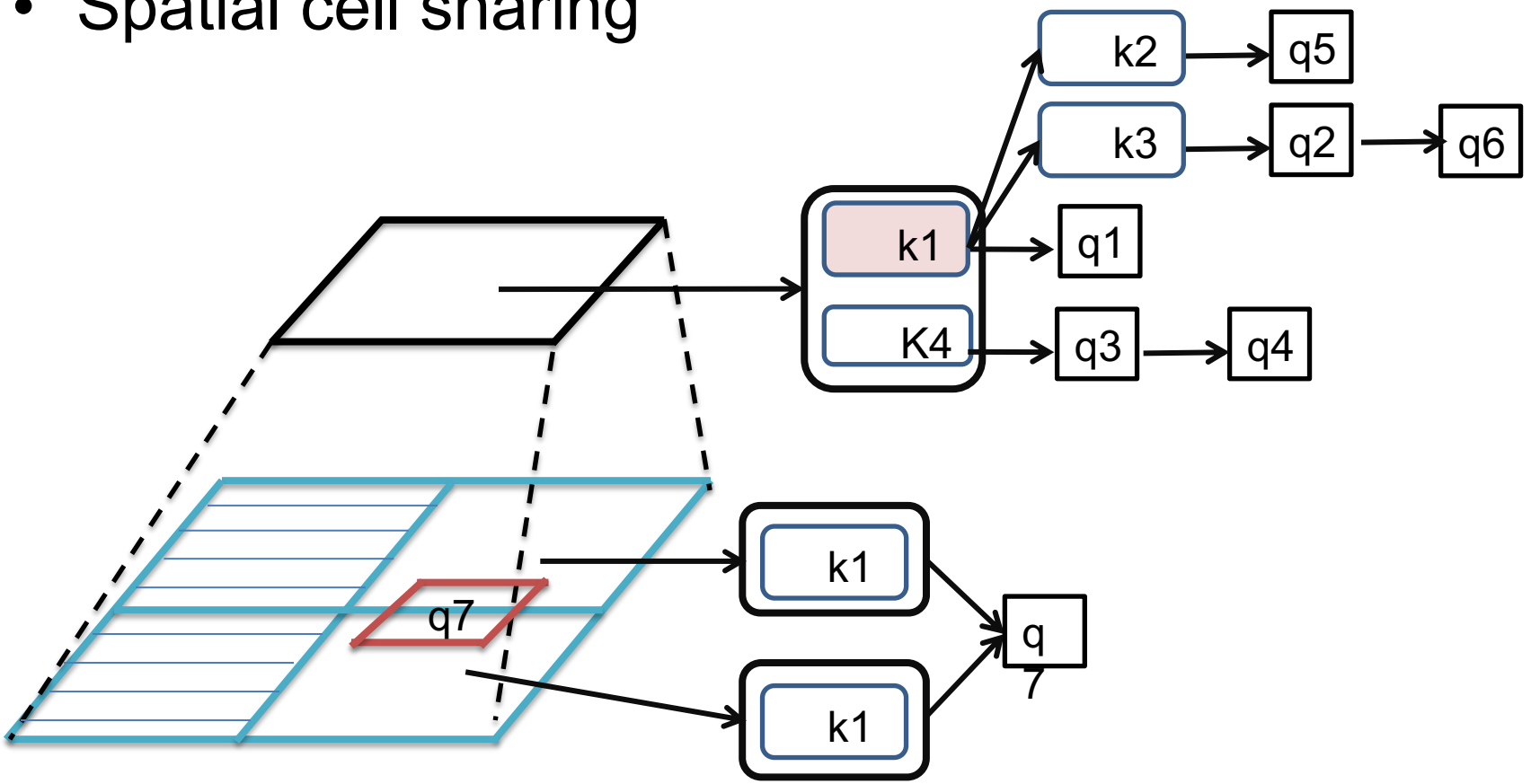
- Integrate the spatial pyramid



Replication that increases the memory overhead

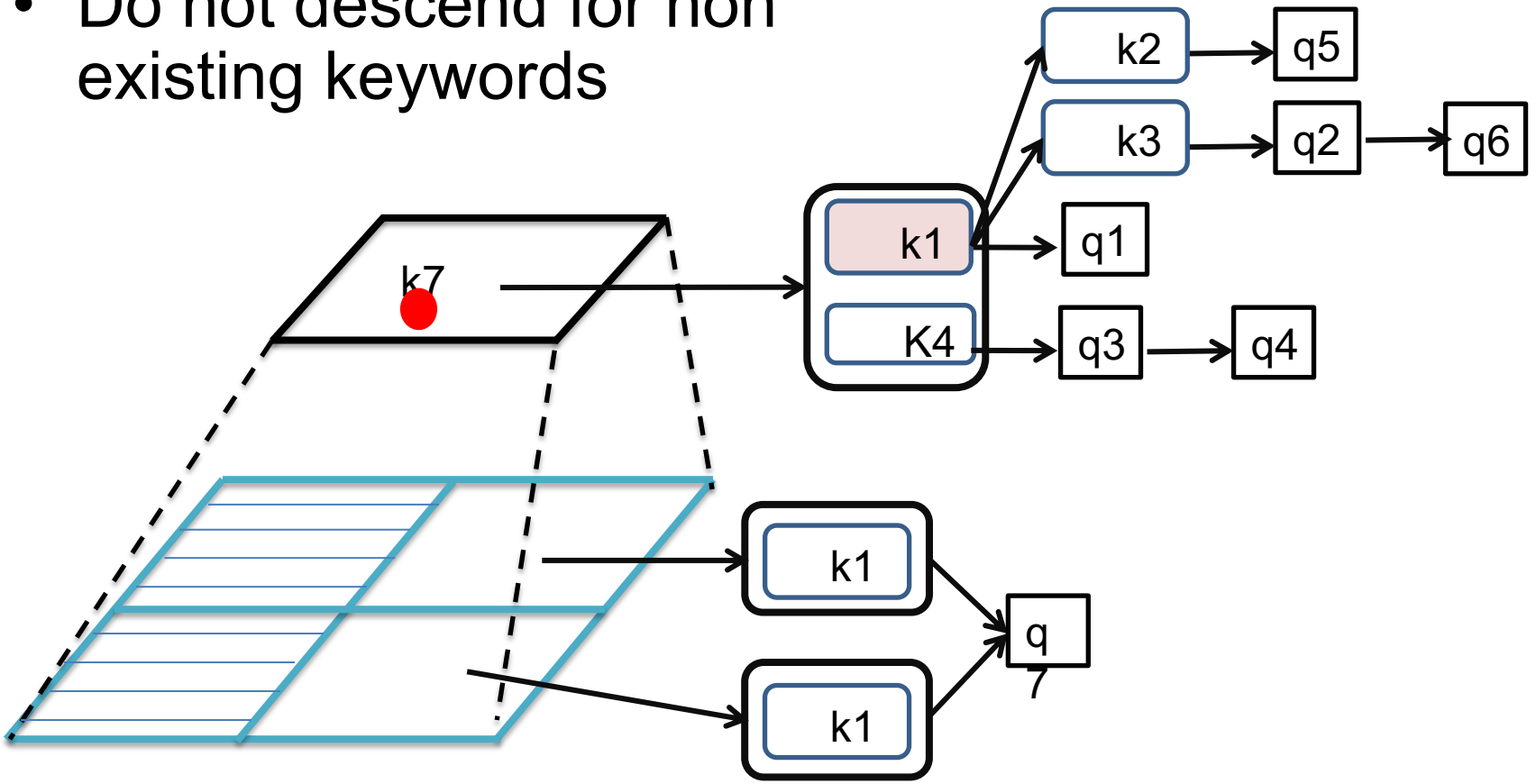
Memory Optimization in FAST

- Spatial cell sharing



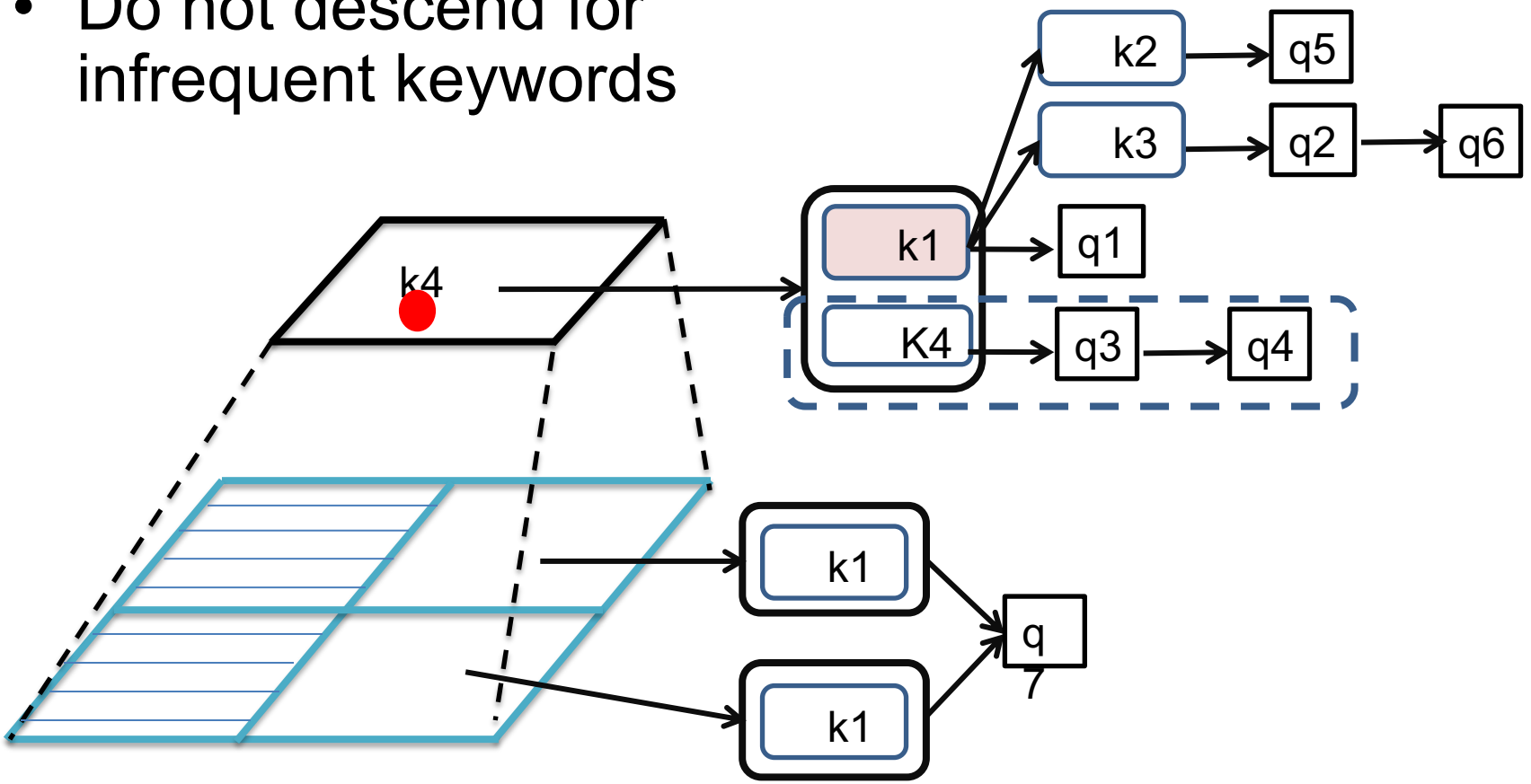
Object Matching: Searching FAST

- Do not descend for non existing keywords



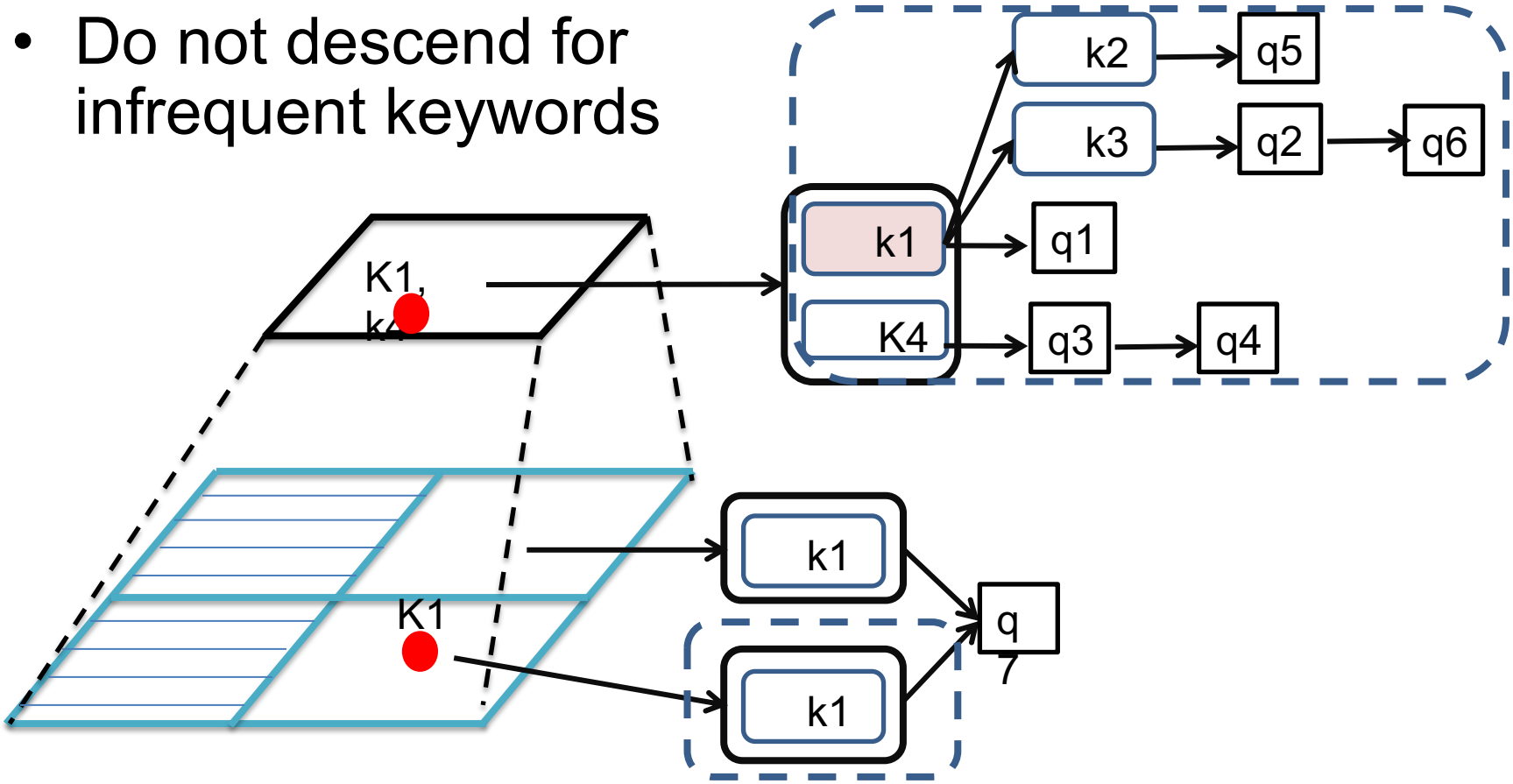
Object Matching: Searching FAST

- Do not descend for infrequent keywords



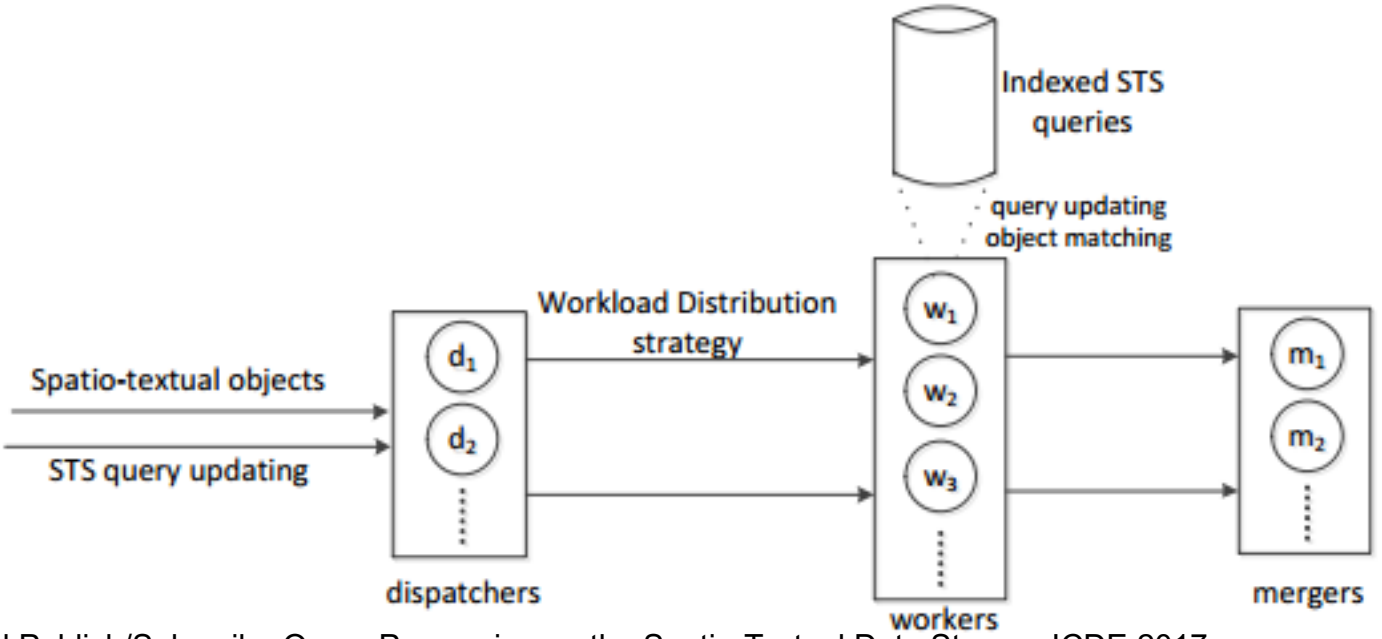
Object Matching: Searching FAST

- Do not descend for infrequent keywords



PS²Stream

- Distributed publish/subscribe system
- Extends the storm streaming system



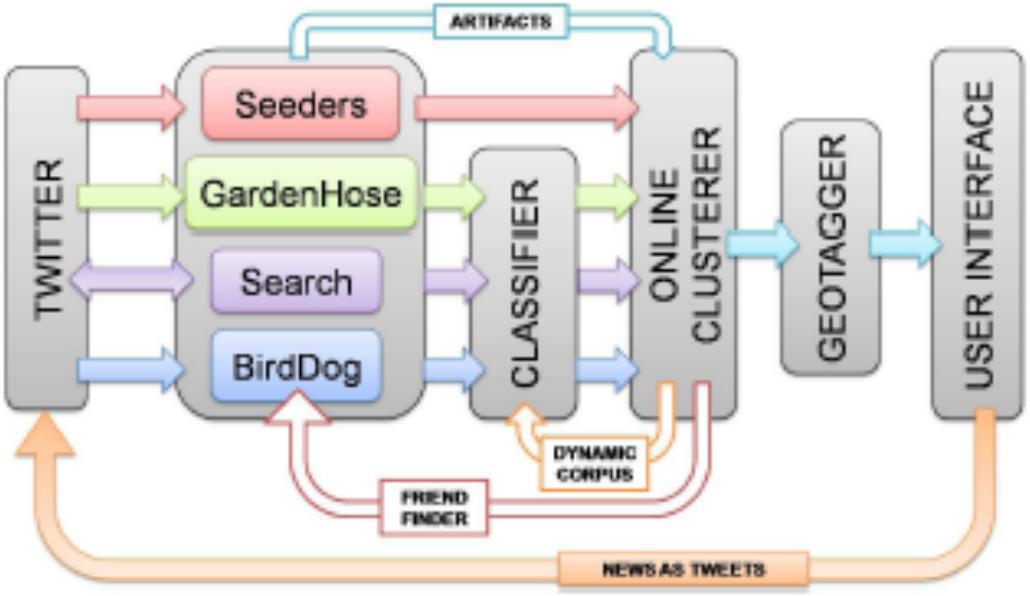
Chen et al., Distributed Publish/Subscribe Query Processing on the Spatio-Textual Data Stream, ICDE 2017

Scalable Spatial-Keyword Systems

- Query-Specific Algorithms
- Extensions to general-purpose big-data systems
- Spatial-keyword-only systems

TwitterStand

- Detection of news from Twitter



Sankaranarayanan et al., Twitterstand: news in tweets, SIGSPATIAL 2009

TwitterStand

Diplomats deliver ultimatum on Honduras coup
Less than 1 hour ago - guardiannews
Diplomats deliver ultimatum on Honduras coup
1011 tweets - Similar Stories - Original Source - Locations

The news is all over Michael Jackson while North Korea is threating us and testing missiles with no news coverage
Less than 1 hour ago - Twitter-Streaming
The news is all over Michael Jackson while North Korea is threating us and testing missiles with no news coverage
1833 tweets - Similar Stories - Original Source - Locations

thinks jacksons dad Joe should not be anywhere near those kids he abused Michael and will do the same to Michaels kids Give Debbie rowe a go
Less than 1 hour ago - Twitter-Streaming
thinks jacksons dad Joe should not be anywhere near those kids he abused Michael and will do the same to Michaels kids Give Debbie rowe a go
574 tweets - Similar Stories - Original Source - Locations

Marines establish positions in Afghan assault
Less than 1 hour ago - timesnews
Marines establish positions in Afghan assault
209 tweets - Similar Stories - Original Source - Locations

62 IranElection Tehran Mousavi Iran neda Neda How to send an anonymous email
1 hours ago - Twitter-Streaming
62 IranElection Tehran Mousavi Iran neda Neda How to send an anonymous email
5422 tweets - Similar Stories - Original Source - Locations

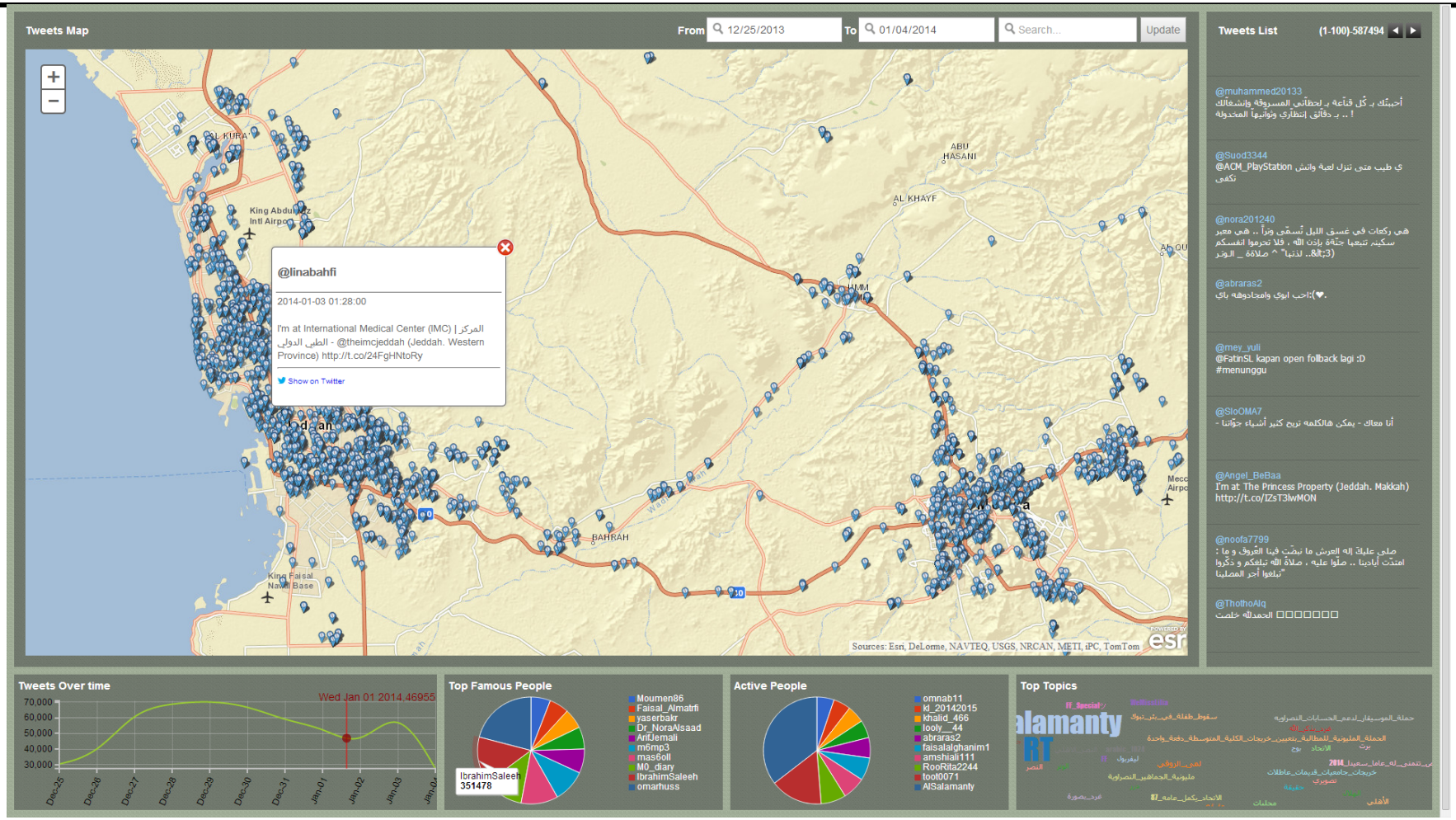
Sanford to reveal schedule details
Less than 1 hour ago - Mediate
Sanford to reveal schedule details
1175 tweets - Similar Stories - Original Source - Locations

TAGHREED

- Answers spatial-keyword queries over micro-blogs including:
  - Spatial-keyword filter
  - Top-k frequent keywords
  - Top-k active users
  - Top-k famous users
  - Aggregates over micro-blogs
  - Top-k languages used while tweeting

Magdy et al., Taghreed: A System for Querying, Analyzing, and Visualizing Geotagged Microblogs, SIGSPATIAL 2014

TAGHREED


Tweets Map
From 12/25/2013
To 01/04/2014
Search...
Update
Tweets List
(1-100)-587494
@linabahfi
2014-01-03 01:28:00
I'm at International Medical Center (IMC) | المركز الطبي الدولي - @theimcjeddah (Jeddah, Western Province) http://t.co/24FgHNtoRy
Show on Twitter
@muhammed20133
أحبِبْك بـ كُل قناعة بـ لحظآتي المسروقة وإنشغآلك بـ .. بـ دقائق إنتظآري وتوآنيها المحدولة !
@Suod3344
@ACM_PlayStation ي طيب متى تنزل لعبة واتش تكفى
@nora201240
هي ركعات في غسق الليل تُسقى وتراً .. هي معبر سكينه تتبعها جنّة بإذن الله ، فلا تحرموا انفسكم لذتها _ صلاة ^ الوتر <3)..
@abraras2
احب ابوي وامجادوهه باي:(♥.
@mey_yuli
@FatinSL kapan open follback lagi :D #menunggu
@SloOMA7
أنا معاك - يمكن هالكلمه تربح كثير أشياء جوّانا
@Angel_BeBaa
I'm at The Princess Property (Jeddah, Makkah) http://t.co/IZsT3IwMON
@noofa7799
صلى عليكَ إله العرش ما نبضَت فينا العُروق و ما : امتدّت أيادينا .. صلّوا عليه ، صلاة الله تبلغكم و ذكّروا تبلغوا أجر المصلينا"
@ThothoAlq
الحمدلله خلصت □□□□□□□
Tweets Over time
Wed Jan 01 2014 46955
70,000
60,000
50,000
40,000
30,000
Dec-25
Dec-26
Dec-27
Dec-28
Dec-29
Dec-30
Dec-31
Jan-01
Jan-02
Jan-03
Top Famous People
IbrahimSaleeh
351478
Moumen86
Faisal_Almatrfi
yaserbakr
Dr_NoraAlsaad
AriJemali
m6mp3
mas6oll
M0_diary
IbrahimSaleeh
omarhuss
Active People
omnab11
kl_20142015
khalid_466
looly__44
abraras2
faisalalghanim1
amshialli111
RooRita2244
tool0071
AlSalamanty
Top Topics
Sources: Esri, DeLorme, NAVTEQ, USGS, NRCAN, METI, iPC, TomTom

Outline

- Introduction and Background
- Querying Spatial-Keyword Data
- Scalable Spatial-Keyword Systems
- Research Directions

Research Directions

- Benchmarking
- Pipelined Evaluation
- Big Spatio-Temporal-keyword Processing

Benchmarking Big Spatial-Keyword Systems

- Spatial-only benchmarks
  - Jackpine
  - BerlinMOD
- Relational benchmarks
  - TPC-H

http://csng.cs.toronto.edu/projects/20
http://dna.fernuni-hagen.de/secondo/BerlinMOD/BerlinMOD.html

Benchmarking Big Spatial-Keyword Systems

- Existing spatial-keyword benchmarks
  - Limited set of queries
  - Specific use cases, e.g., social-network analysis

Chen et al., Spatial keyword query processing: An experimental evaluation, PVLDB 2013
Panda et al., Performance Evaluation of Social Network Using Data Mining Techniques, Computational Social Networks 2012

Benchmarking Big Spatial-Keyword Systems

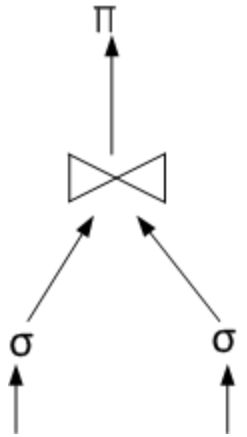
- Large spatial-keyword datasets
- Support of a wide range of spatial-keyword queries
- Support of several evaluation scenarios
  - Batching
  - Streaming

Research Directions

- Benchmarking
- Pipelined Evaluation
- Big Spatio-Temporal-keyword Processing

Pipelined Evaluation of Big Spatial-keyword Queries

- Design building-block operators that can be composed to realize complex spatial-keyword queries
  - Similar to the relational SELECT, PROJECT, JOIN building-block operators



Research Directions

- Benchmarking
- Pipelined Evaluation
- Big Spatio-Temporal-keyword Processing

Big Spatio-Temporal-keyword Processing

- Spatial-keyword data are often associated with temporal attributes
  - Time-stamp of tweets
- We need to investigate
  - Indexing
  - Processing
  - Query processing and Query languages

Thank you!

The authors acknowledge the support of the National Science Foundation under Grant Numbers III-1815796 and IIS-1910216.