The Spatial Web – A New Data Management Frontier

Christian S. Jensen

www.cs.au.dk/~csj

AARHUS UNIVERSITY

The Web Is Going Mobile

- A quickly evolving mobile Internet infrastructure.
  - Mobile devices, e.g., smartphones, tablets, laptops, navigation devices, glasses
  - Communication networks and users with access
- Sales
  - Smartphones: 2010: 310 million: 2011: 490 million; 2012: 650-690 million; 2016: 1+ billion (half of the phone market)
  - PCs (desktop, laptop): 2010: 350 million; 2011: 350 million
  - Tablets: 2011: 66 million
- Going Mobile is a mega trend.
  - Google went “mobile first” in 2010.
  - Mobile data traffic 2020 = 2010 x 1000.

Mobile Is Spatial

- Increasingly sophisticated technologies enable the accurate geo-positioning of mobile users.
  - GPS-based technologies
  - Positioning based on Wi-Fi and other communication networks
  - New technologies are underway (e.g., GNSSs and indoor).

Outline

- Mobile location-based services
- Spatial keyword querying
  - Top-*k* spatial keyword queries
  - Continuous top-*k* queries
  - Accounting for co-location
  - Collective queries
- Place ranking using user-generated content
  - GPS records, directions queries
- Summary and challenges

(Acknowledgments and references are given at the end:
see also the paper in the proceedings.)

Transportation-Related Services

- Spatial pay per use, or metered services
  - E.g., road pricing: payment based on where, when, and how much one drives; insurance; parking
- Eco routing and driving
  - Reduction of GHG emissions, an important element in combating global warming (e.g., [reduction-project.eu])
- Self-driving vehicles
  - “…looking back and saying how ridiculous it was that humans were driving cars.” [Sebastian Thrun, TED2011]
  - Machines don’t make mistakes, human do.

Location-Based Games

- Move games from going on behind a computer or phone display to occur reality.
- Virtual objects, seen by the players on their displays, are given physical locations that are know to the system.
- Physical objects, the players, are being tracked by the system.

- Virtual playgrounds for kids (e.g., [playingmondo.com])
- Paintball (e.g., Botfighters 2.0)
- “Catch the monsters” (e.g., Raygun)

[IEEE Spectrum 43(1), Jan 2006]

Spatial Web Querying

- Total web queries
  - Google: 2011 daily average: 4.7 billion
- Queries with local intent
  - ”cheap pizza” vs. ”pizza recipe”
  - Google: ~20% of desktop queries
  - Bing: 50+% of mobile queries
- Vision: Improve web querying by exploiting accurate user and content geo-location
  - Smartphone users issue keyword-based queries
  - The queries concern websites for places
- Balance spatial proximity and textual relevance

***Top-k* spatial keyword querying**

Top-*k* Spatial Keyword Query

- Objects: $p = \langle \lambda, \psi \rangle$ (location, text description)
- Query: $q = \langle \lambda, \psi, k \rangle$ (location, keywords, # of objects)

- Ranking function

$$rank_q(p) = \alpha \frac{\| q.\lambda, p.\lambda \|}{\max D} + (1-\alpha)(1 - \frac{tr_{q.\psi}(p.\psi)}{\max P}) \qquad 0 \le \alpha \le 1$$

  - Distance: $\| q.\lambda, p.\lambda \|$
  - Text relevancy: $tr_{q.\psi}(p.\psi)$
    - Probability of generating the keywords in the query from the language models of the documents

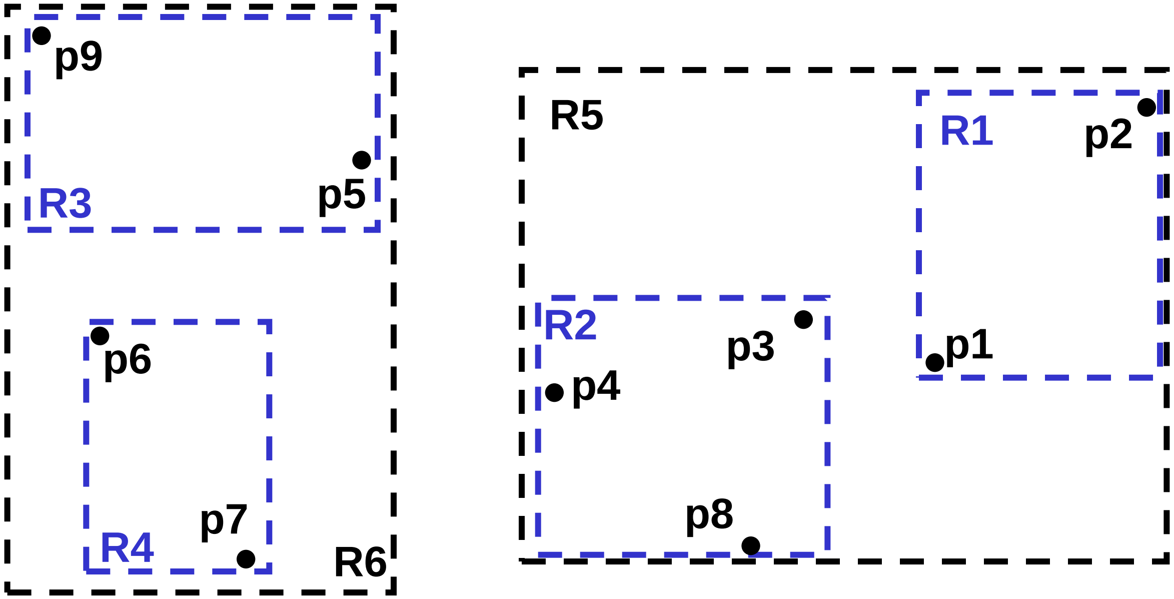
- Generalizes the *k*NN query and text retrieval

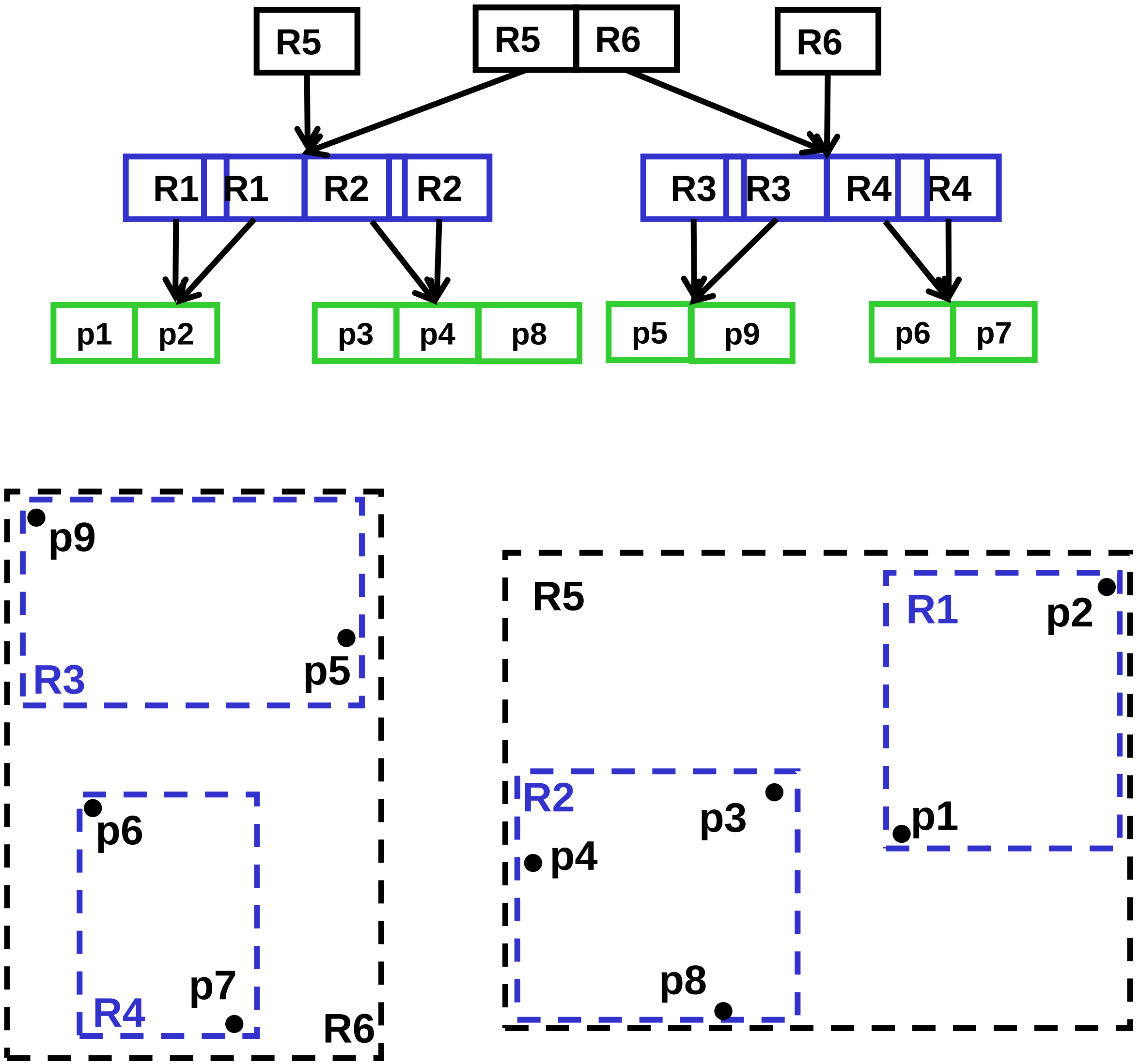
Spatial Keyword Query Processing

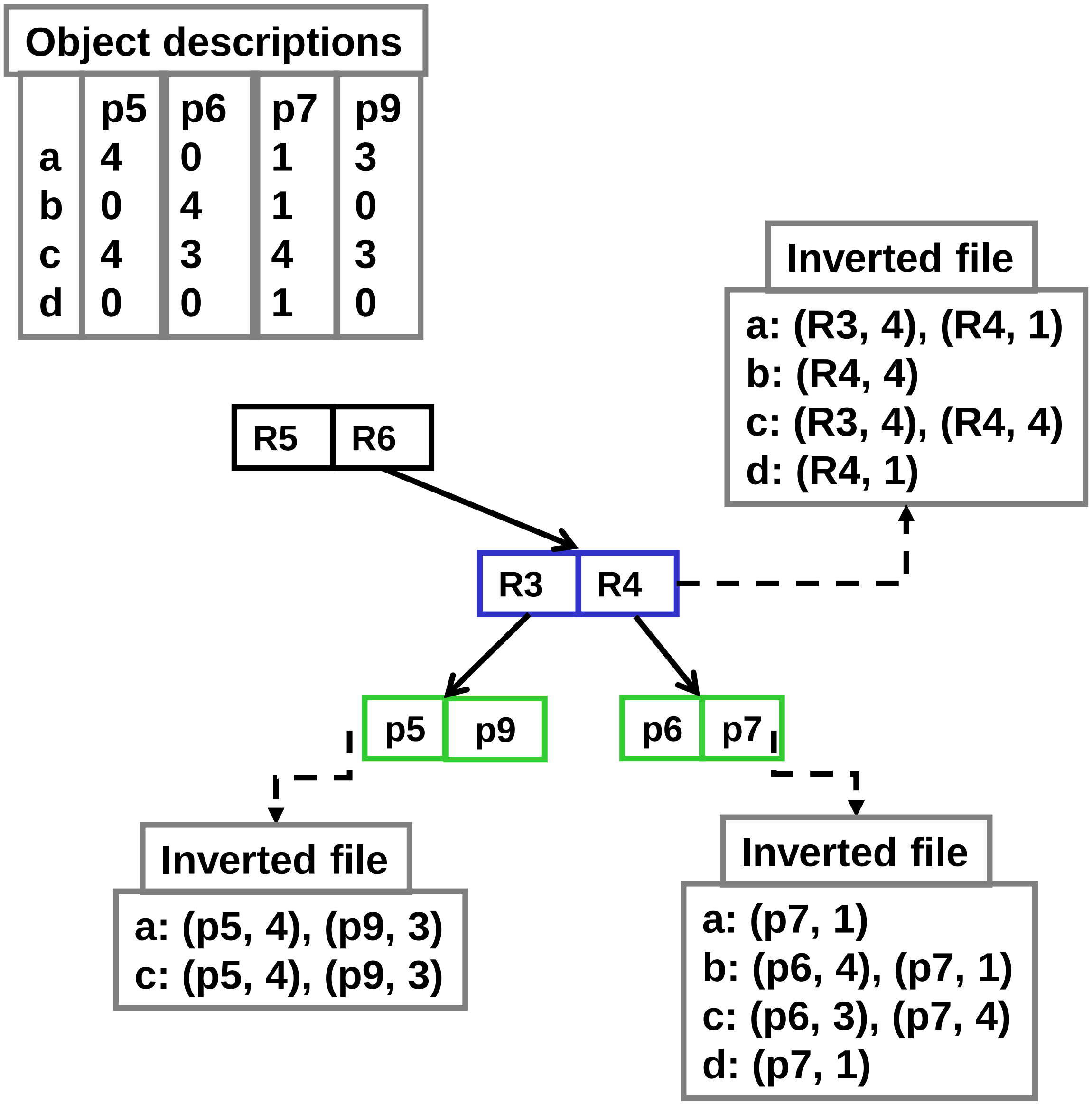
- How do we process spatial keyword queries efficiently?

- Proposal
  - Prune both spatially and textually in an integrated fashion
  - Apply indexing to accomplish this

- The IR-tree [Cong et al. 2009 ; Li et al. 2011]
  - Combines the R-tree with inverted files
  - R-tree: good for spatial
  - Inverted files: good for text


p9
R3
p5
p6
R4
p7
R6
R5
R1
p2
R2
p3
p1
p4
p8


R5
R5
R6
R6
R1
R1
R2
R2
R3
R3
R4
R4
p1
p2
p3
p4
p8
p5
p9
p6
p7
p9
R3
p5
p6
R4
p7
R6
R5
R1
p2
R2
p3
p1
p4
p8



Object descriptions

	p5	p6	p7	p9
a	4	0	1	3
b	0	4	1	0
c	4	3	4	3
d	0	0	1	0

Inverted file

a: (R3, 4), (R4, 1)
b: (R4, 4)
c: (R3, 4), (R4, 4)
d: (R4, 1)

Inverted file

a: (p5, 4), (p9, 3)
c: (p5, 4), (p9, 3)

Inverted file

a: (p7, 1)
b: (p6, 4), (p7, 1)
c: (p6, 3), (p7, 4)
d: (p7, 1)

Continuous top-*k* querying

Continuous Spatial Keyword Queries

- Objects: $p = \langle \lambda, \psi \rangle$ (location and text description)
- Query: $q = \langle \lambda, \psi, k \rangle$ (location, keywords, # of objects)
- A continuous query where argument λ changes continuously

- Ranking function

$$rank_q(p) = \frac{\| q.\lambda, p.\lambda \|}{tr_{q.\psi}(p.\psi)}$$

Euclidean distance (changes continuously)

Text relevancy (query dependent)

Continuous Spatial Keyword Queries

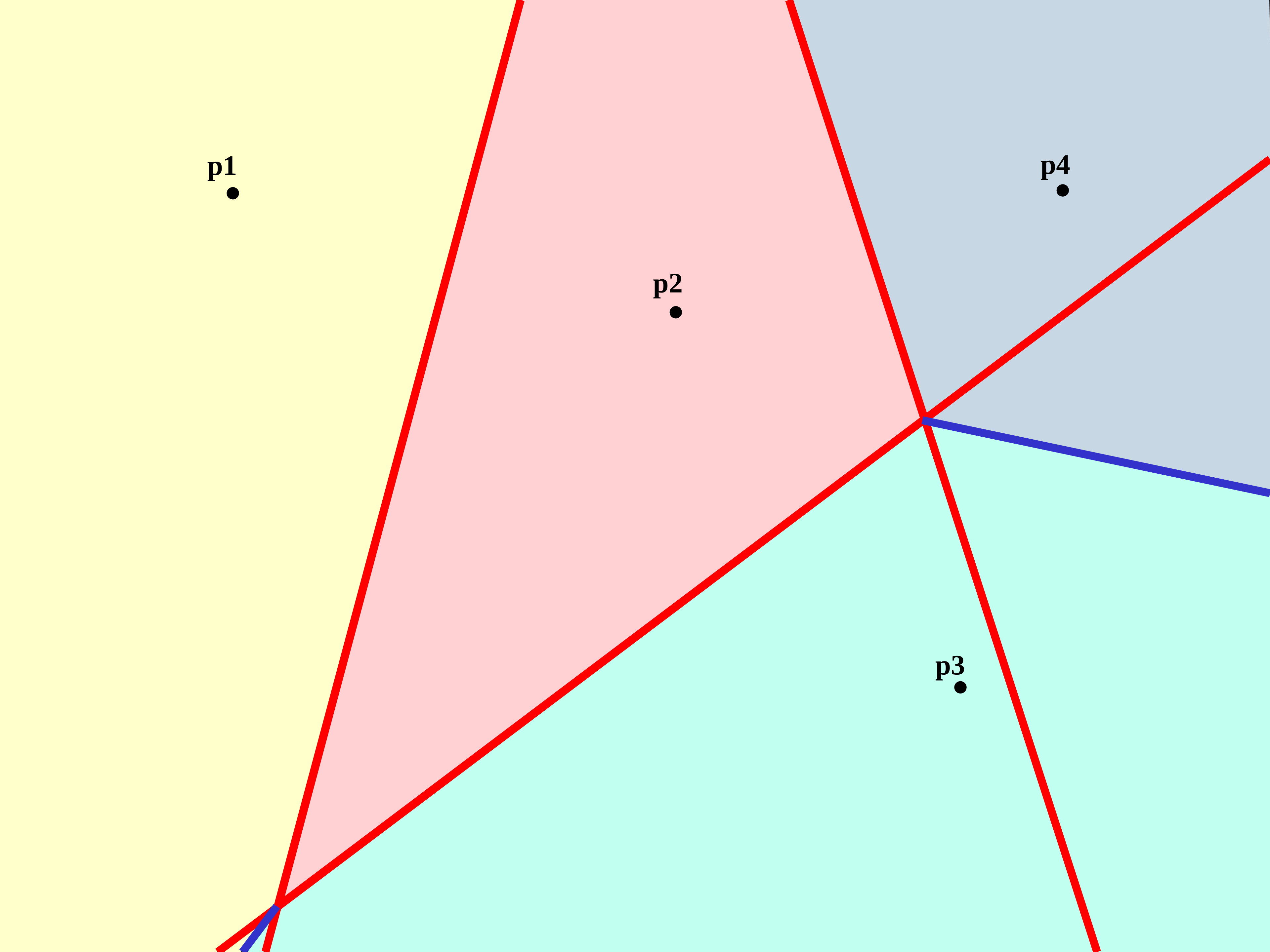
- How can we process such queries efficiently?
  - Server-side computation cost
  - Client-server communication cost

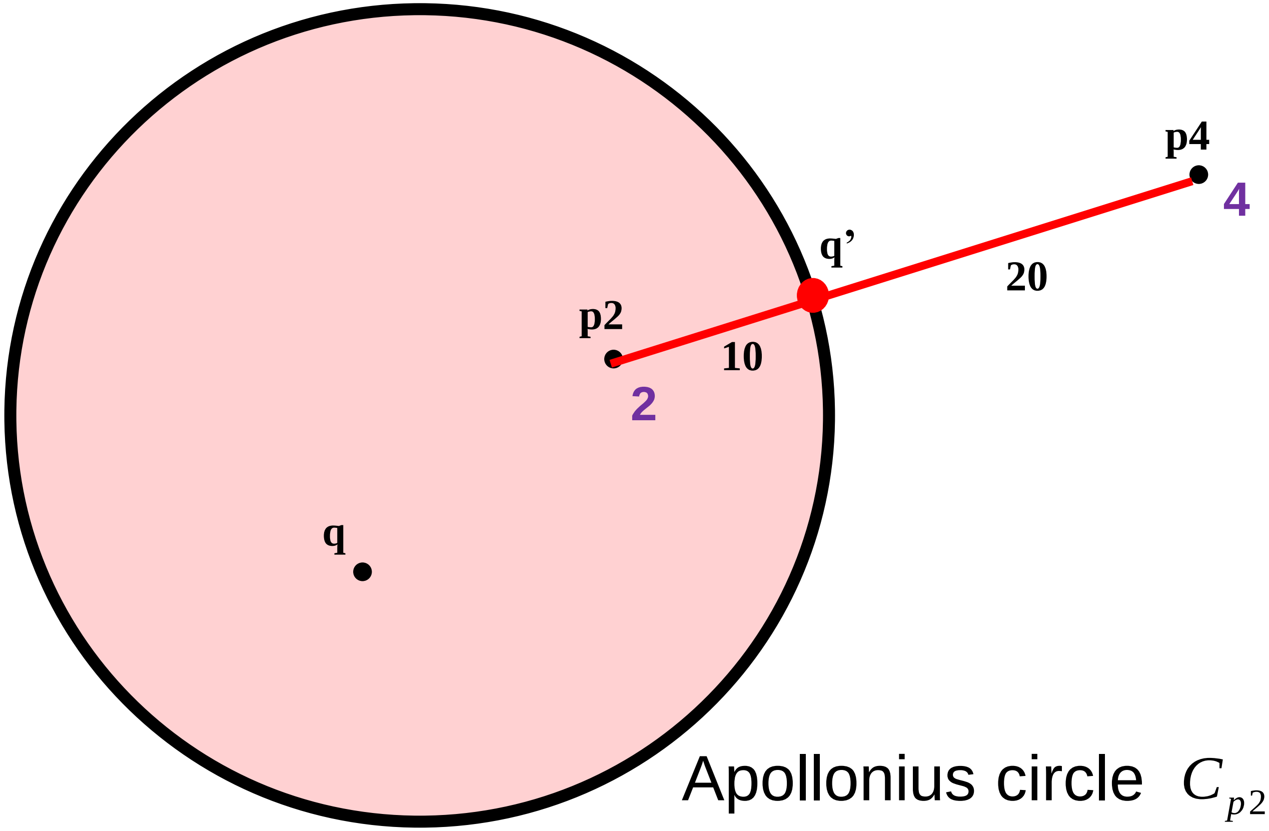
- While the argument changes continuously, the result changes only discretely.
  - Do computation only when the result may have changed

- Use safe zones
  - When the user remains within the zone, the result does not change.
  - The user requests a new result when about to exit the safe zone.

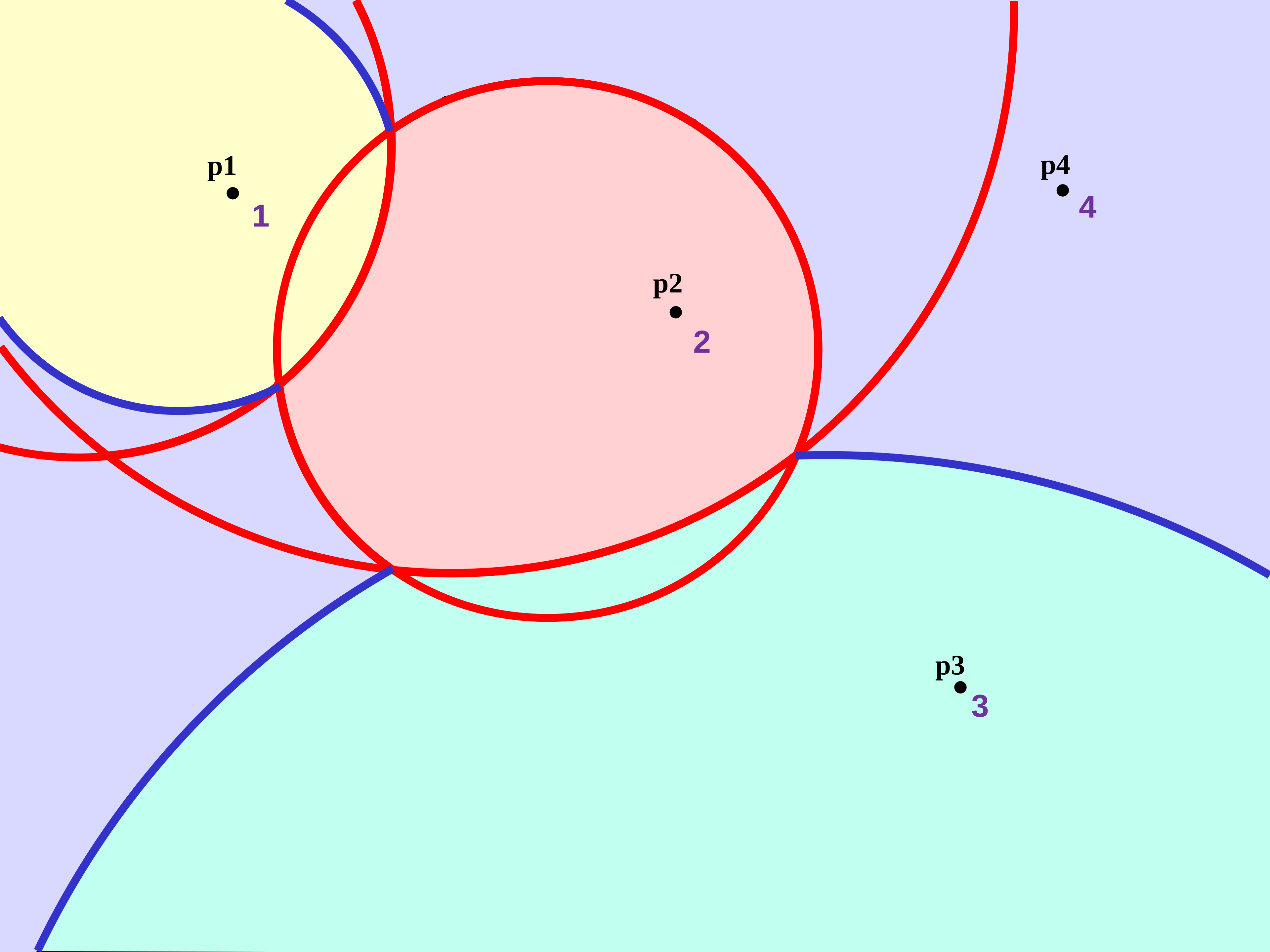
Processing Continuous Queries

- Compute results
  - As before…
- Compute corresponding safe zones
  - Integrate with result computation
- Prune objects that do not contribute to the safe zone without inspecting them
  - Use the IR-tree
  - Access objects in border-distance order
  - Prune sub-trees
  - Terminate safely when a stopping criterion is met


p1
p2
p4
p3



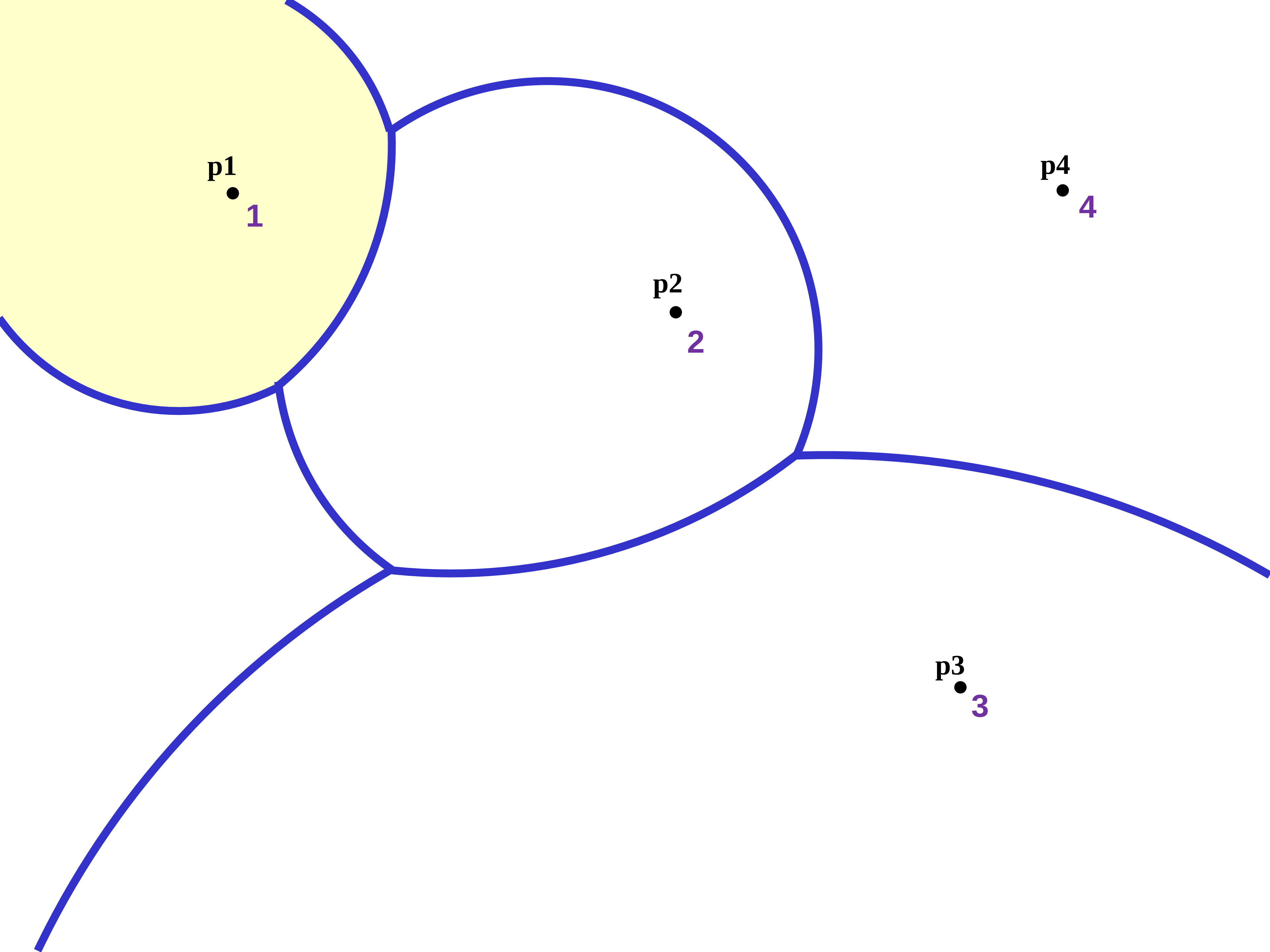
Apollonius circle $C_{p2,p4}$


p1
1
p2
2
p4
4
p3
3

Representation of a Multiplicatively Weighted Voronoi Cell

Influence Objects

$$I^{+} \cup I^{o} \cup I^{-}$$


p1
1
p2
2
p4
4
p3
3

Pruning Objects p_+ with Higher Weights

$$\exists p' \in I^+ (C_{p^*, p_+} \supseteq C_{p^*, p'})$$

Pruning Objects with Equal Weights

$$\exists p' \in I^+ (\perp_{p^*, p_o} \supseteq C_{p^*, p'})$$

$$\exists p' \in I^o (\perp_{p^*, p_o} \supseteq \perp_{p^*, p'})$$

Pruning Objects with Lower Weights

$$\exists p' \in I^+ (C_{p_-, p^*} \cap C_{p^*, p'} = \emptyset)$$

$$\exists p' \in I^- (C_{p_-, p^*} \subseteq C_{p', p^*})$$

$$\exists p' \in I^o (C_{p_-, p^*} \cap \perp_{p^*, p'} = \emptyset)$$

Prestige-based ranking

Accounting for Co-Location

- So far, we have considered data objects as independent, but they are not.

- It is common that similar places co-locate.
  - Markets with many similar stands
  - Shopping centers, districts
  - China town, little India, little Italy, …
  - Restaurant and bar districts
  - Car dealerships

- How can we capture and take into account the apparent benefits of co-location?

Top-*k* Spatial Keyword Query

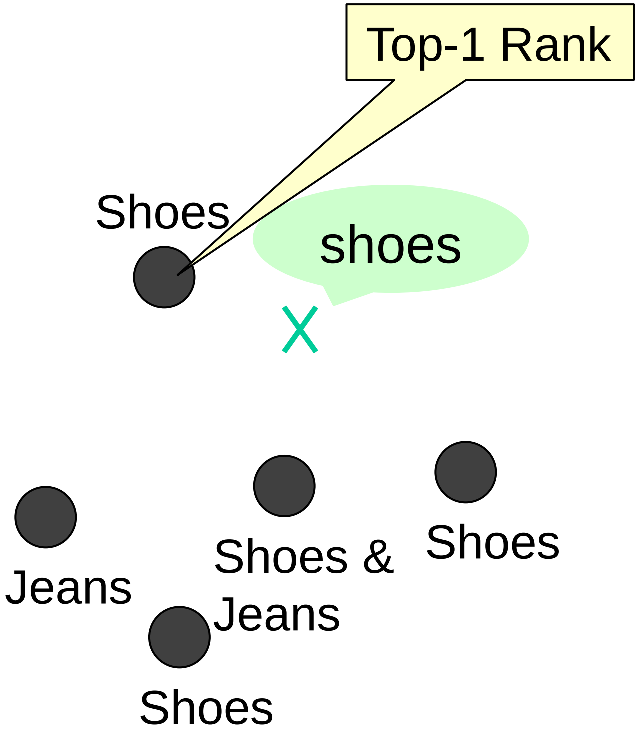
- Objects: $p = \langle \lambda, \psi \rangle$ (location, text description)
- Query: $q = \langle \lambda, \psi, k \rangle$ (location, keywords, # of objects)

- Ranking function

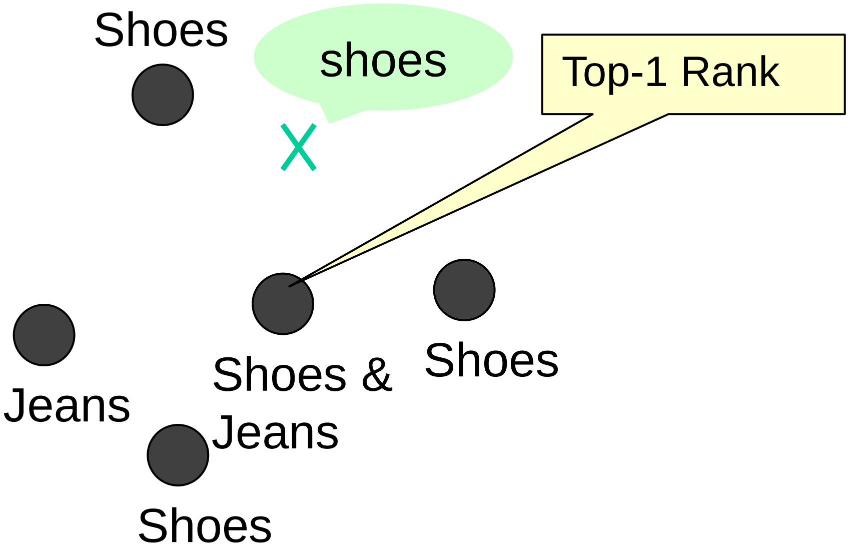
$$prrank_q(p) = \alpha \frac{\| q.\lambda, p.\lambda \|}{\max D} + (1-\alpha)(1 - pr_{q.\psi}(p.\psi)) \qquad 0 \le \alpha \le 1$$

  - Distance: $\| q.\lambda, p.\lambda \|$
  - Text relevancy: $pr_{q.\psi}(p.\psi)$
    - PR score: prestige-based text relevancy (normalized)

First Retrieval Approach



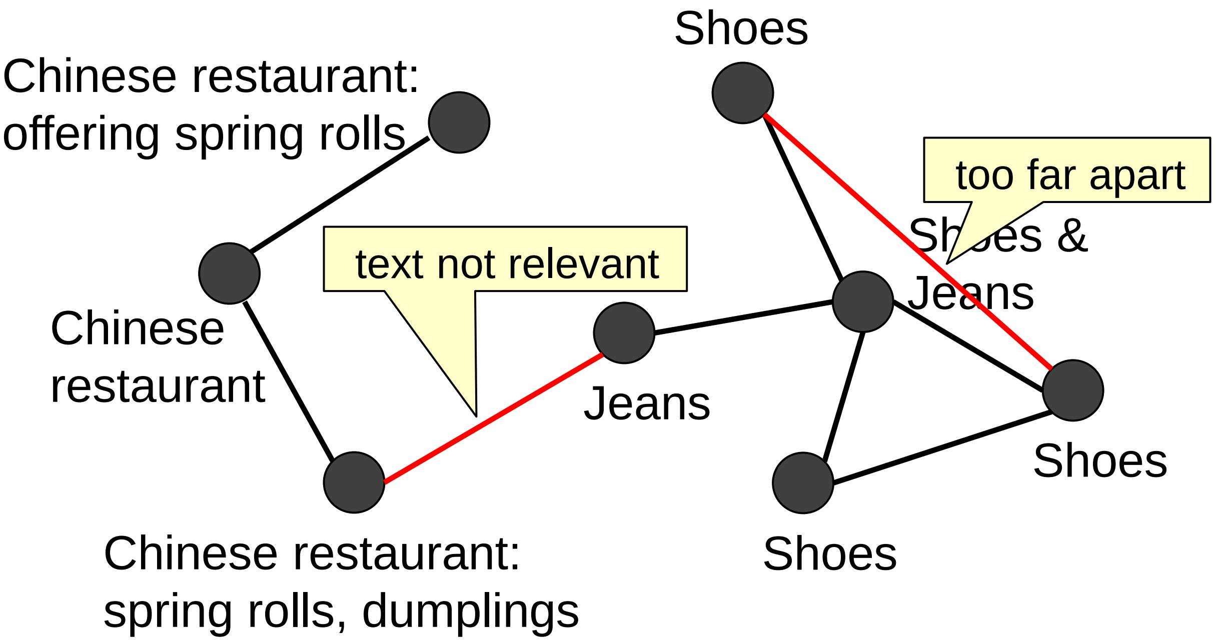
Prestige-Based Retrieval



Prestige-Based Ranking

- Prestige propagation using a graph **G** = (V, E, W)
  - Vertices V: spatial web objects
  - Edges E: connect objects that meet constraints
  - Distance threshold: $\| p_i.\lambda, p_j.\lambda \| \leq \lambda$
  - Similarity threshold: $sim(p_i.\psi, p_j.\psi) \leq \xi$ (vector space model)
  - Edge weights W: $\| p_i.\lambda, p_j.\lambda \|$
- Use Personalized PageRank for ranking [Jeh & Widom, 2003]

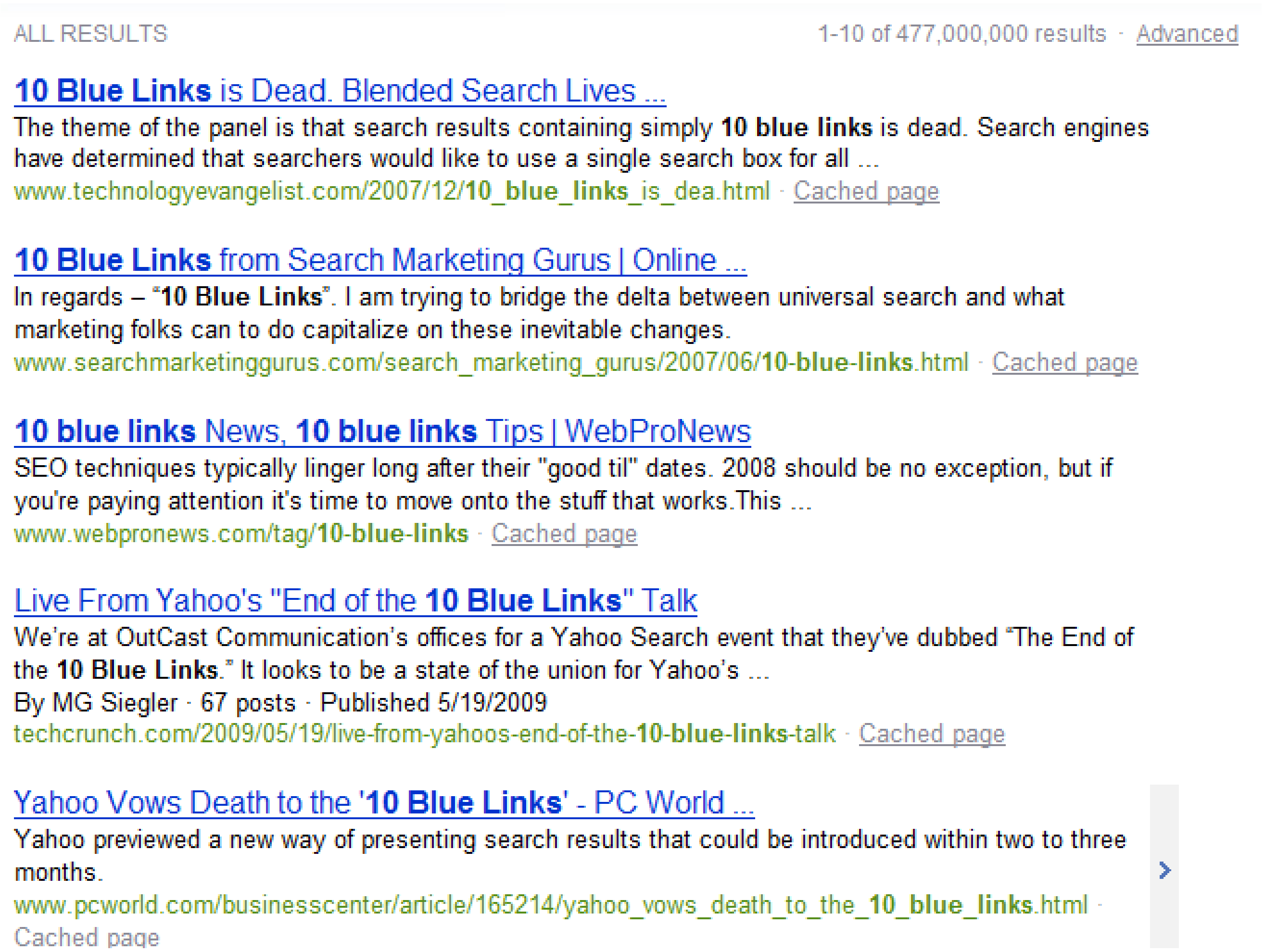
Prestige-Based Ranking



Experimental Study

- Local experts are asked to provide query keywords for locations and then to evaluate the results of the resulting queries.
- The studies suggest that the approach is able to produce better results than is the baseline without score propagation.

Collective queries



ALL RESULTS 1-10 of 477,000,000 results · Advanced

10 Blue Links is Dead. Blended Search Lives ...

The theme of the panel is that search results containing simply **10 blue links** is dead. Search engines have determined that searchers would like to use a single search box for all ...

www.technologyevangelist.com/2007/12/**10_blue_links**_is_dea.html · Cached page

10 Blue Links from Search Marketing Gurus | Online ...

In regards – "**10 Blue Links**". I am trying to bridge the delta between universal search and what marketing folks can to do capitalize on these inevitable changes.

www.searchmarketinggurus.com/search_marketing_gurus/2007/06/**10-blue-links**.html · Cached page

10 blue links News, **10 blue links** Tips | WebProNews

SEO techniques typically linger long after their "good til" dates. 2008 should be no exception, but if you're paying attention it's time to move onto the stuff that works.This ...

www.webpronews.com/tag/**10-blue-links** · Cached page

Live From Yahoo's "End of the **10 Blue Links**" Talk

We're at OutCast Communication's offices for a Yahoo Search event that they've dubbed "The End of the **10 Blue Links**." It looks to be a state of the union for Yahoo's ...

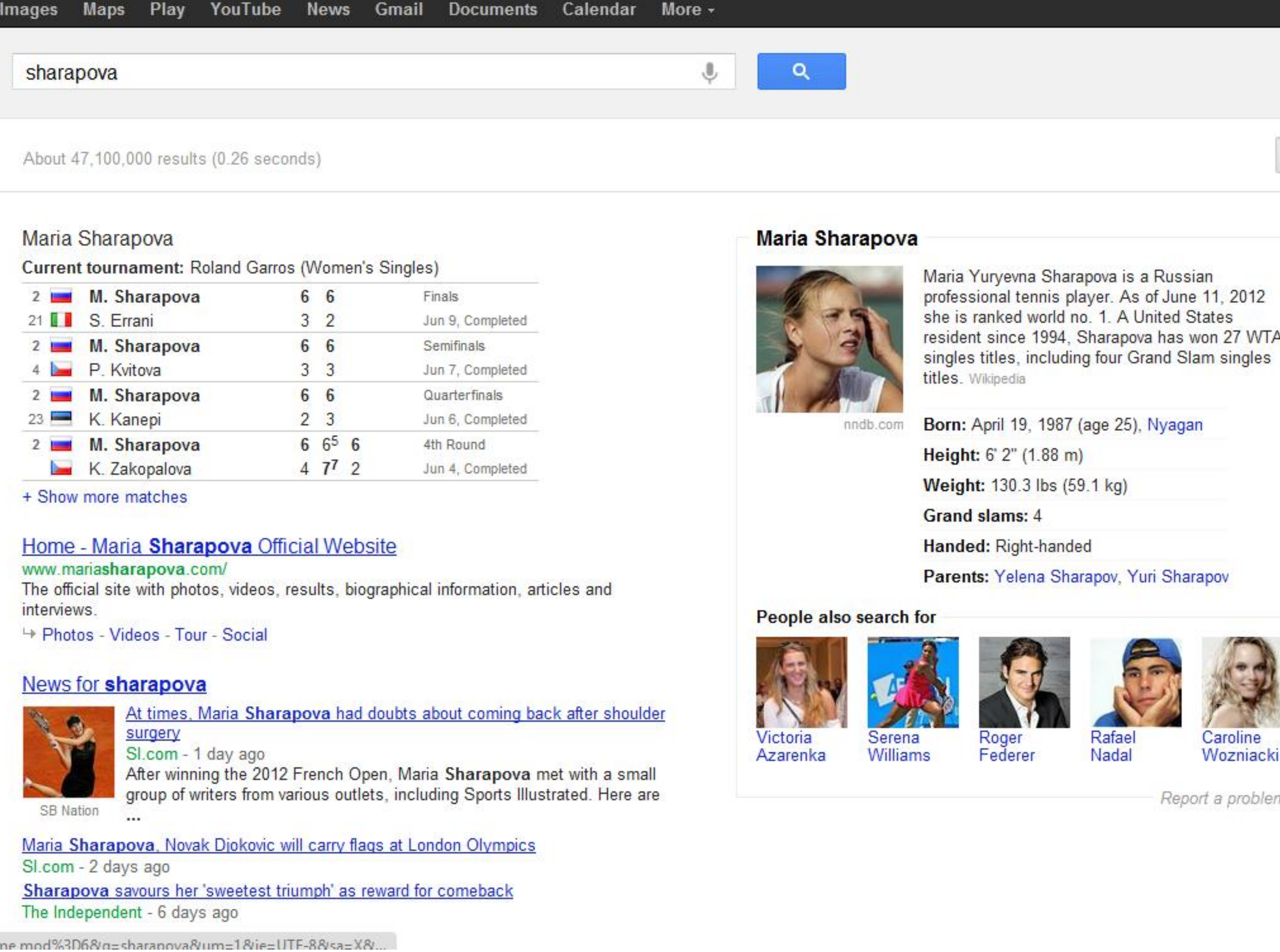
By MG Siegler · 67 posts · Published 5/19/2009

techcrunch.com/2009/05/19/live-from-yahoos-end-of-the-**10-blue-links**-talk · Cached page

Yahoo Vows Death to the '**10 Blue Links**' - PC World ...

Yahoo previewed a new way of presenting search results that could be introduced within two to three months.

www.pcworld.com/businesscenter/article/165214/yahoo_vows_death_to_the_**10_blue_links**.html · Cached page


Images Maps Play YouTube News Gmail Documents Calendar More ▾
sharapova

About 47,100,000 results (0.26 seconds)

Maria Sharapova

Current tournament: Roland Garros (Women's Singles)

2	🇷🇺	**M. Sharapova**	6	6		Finals
21	🇮🇹	S. Errani	3	2		Jun 9, Completed
2	🇷🇺	**M. Sharapova**	6	6		Semifinals
4	🇨🇿	P. Kvitova	3	3		Jun 7, Completed
2	🇷🇺	**M. Sharapova**	6	6		Quarterfinals
23	🇪🇪	K. Kanepi	2	3		Jun 6, Completed
2	🇷🇺	**M. Sharapova**	6	6^5	6	4th Round
	🇨🇿	K. Zakopalova	4	7^7	2	Jun 4, Completed

+ Show more matches

Home - Maria **Sharapova** Official Website
www.maria**sharapova**.com/
The official site with photos, videos, results, biographical information, articles and interviews.
↳ Photos - Videos - Tour - Social

News for **sharapova**

At times, Maria **Sharapova** had doubts about coming back after shoulder surgery
SI.com - 1 day ago
After winning the 2012 French Open, Maria **Sharapova** met with a small group of writers from various outlets, including Sports Illustrated. Here are ...

SB Nation

Maria **Sharapova**, Novak Djokovic will carry flags at London Olympics
SI.com - 2 days ago

Sharapova savours her 'sweetest triumph' as reward for comeback
The Independent - 6 days ago

ne mod%3D6&q=sharapova&um=1&ie=UTF-8&sa=X&...

Maria Sharapova

nndb.com

Maria Yuryevna Sharapova is a Russian professional tennis player. As of June 11, 2012 she is ranked world no. 1. A United States resident since 1994, Sharapova has won 27 WTA singles titles, including four Grand Slam singles titles. Wikipedia

Born: April 19, 1987 (age 25), Nyagan

Height: 6' 2" (1.88 m)

Weight: 130.3 lbs (59.1 kg)

Grand slams: 4

Handed: Right-handed

Parents: Yelena Sharapov, Yuri Sharapov

People also search for

Victoria Azarenka

Serena Williams

Roger Federer

Rafael Nadal

Caroline Wozniacki

Report a problem

Collective Spatial Keyword Querying

- So far, the granularity of a result has been a single object

- The spatial aspect offers natural ways of aggregating data objects and providing aggregate query results.

- We may want to return *sets* of objects that collectively satisfy a query.

The Spatial Group Keyword Query

- Objects: $o = \langle \lambda, \psi \rangle$ (location and text description)
- Query: $Q = \langle \lambda, \psi \rangle$ (location and keywords)

- The result is a group of objects χ satisfying two conditions.
  - $Q.\psi \subseteq \bigcup_{o \in \chi} o.\psi$
  - $Cost(Q, \chi)$ is minimized.

- $Cost(Q, \chi) = \alpha C_1(Q, \chi) + (1 - \alpha) C_2(\chi)$
  - $C_1(.,.)$ depends on the distances of the objects in χ to Q.
  - $C_2(.)$ characterizes the inter-object distances among objects in χ.
  - α balances the weights of the two components.

Spatial Group Query Variants

- Cost function: $Cost(Q,\chi) = \sum_{o \in \chi} Dist(o,Q)$
- Application scenario
  - The user wishes to visit the places one by one while returning to the query location in-between.
  - Go to the hotel between the museum visit and the jazz concert
  - NP-complete: proof by reduction from the Weighted Set Cover problem

- Cost function: $Cost(Q,\chi) = \max_{o \in \chi} Dist(o,Q) + \max_{o_i, o_j \in \chi} Dist(o_i, o_j)$
- Application scenario
  - Visit places without returning to the query location in-between
  - E.g., go to a movie and then dinner
  - NP-complete: proof from reduction from the 3-SAT problem

skip

Place ranking using GPS records, directions queries

GPS-Based Place Ranking I

- Massive volumes of location samples from moving objects are becoming available.
  - GPS location records *(oid, x, y, t)*
  - Location records based on Wi-Fi and cellular positioning
- How can we utilize this content for ranking spatial web objects?

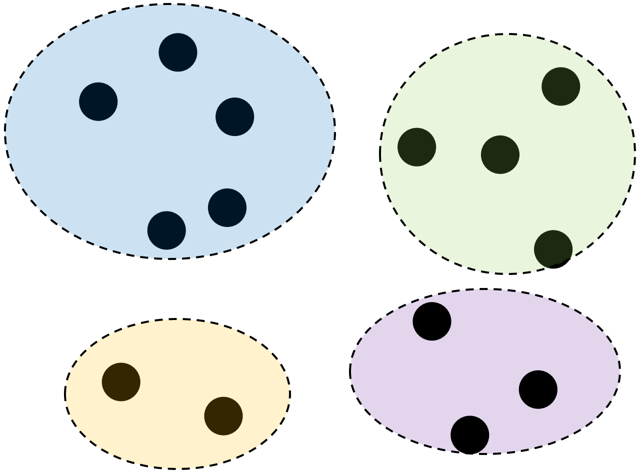
GPS-Based Place Ranking II

- Methodology
  - Connect the GPS data with places (semantic locations)
  - Use the GPS data for ranking the places

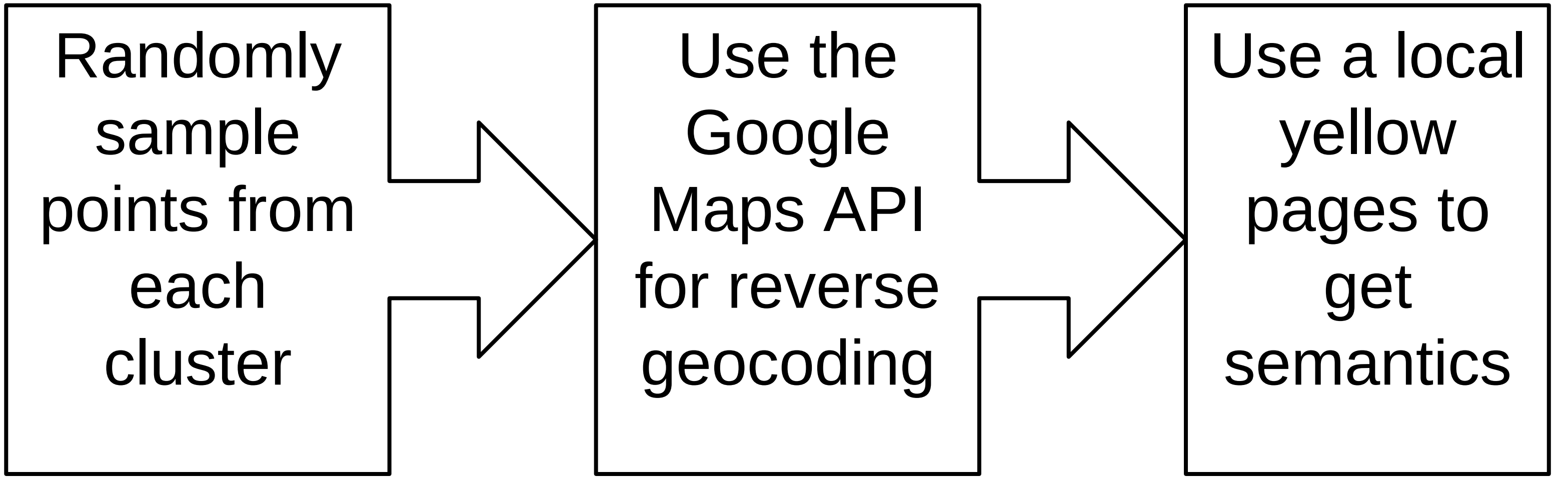
- …in more detail
  - Step 1: Extract stay points from raw trajectories
  - Step 2: Cluster stay points with existing algorithms
  - Step 3: Reverse geocode the stay points and obtain their semantics from business directories
  - Step 4: Refine the clusters to obtain semantic locations
  - Step 5: Ranking

Step 2: Cluster Stay Points

- Use existing spatial clustering algorithms
  - K-means, OPTICS



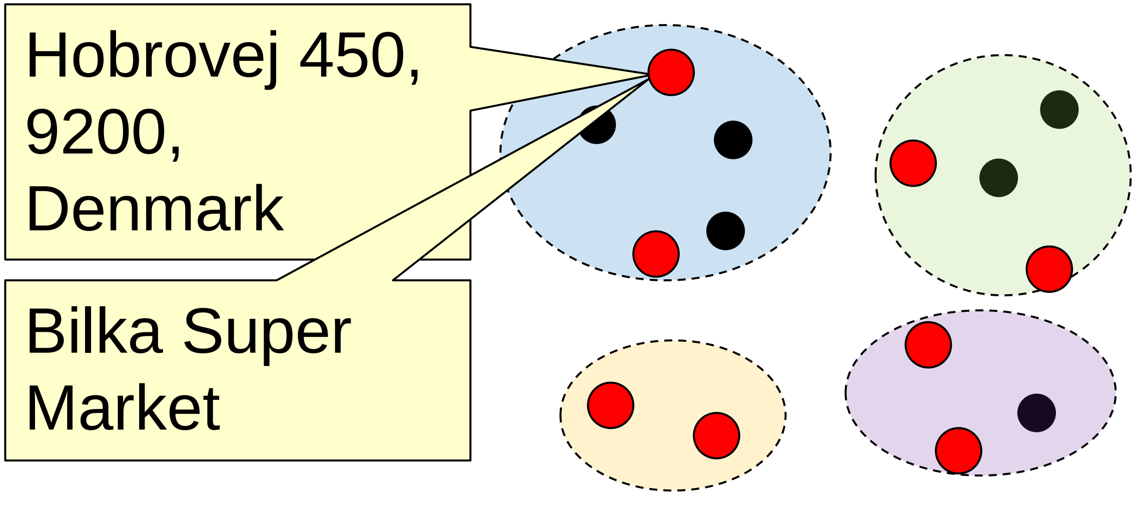
Step 3: Sampling, Reverse Geocoding, Semantics



Randomly sample points from each cluster

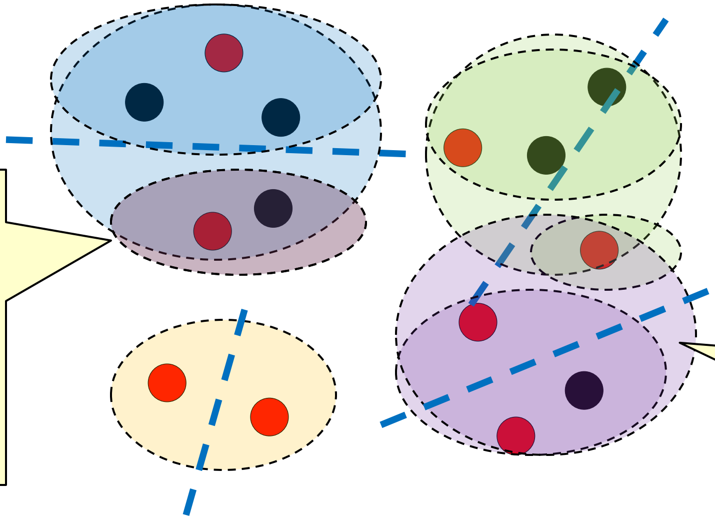
Use the Google Maps API for reverse geocoding

Use a local yellow pages to get semantics



Step 4: Splitting and Merging

- Splitting
  - Cluster points in a cluster to obtain sub-clusters
  - Split a cluster if it has sub-clusters with different semantics
- Merge two clusters with similarity larger than a threshold
  - Similarity: consider user lists, semantics lists, average entry times, average stay durations



Cannot merge with others; becomes a new cluster

These merge to form a new cluster

Experimental Study

- Data
  - Collected from device installed in cars in Nordjylland, Denmark
  - 119 users in the period 01/01/2007 ~ 31/03/2008
  - Sampling @ 1Hz
  - 105,329,114 records
- Step 1 – stay point extraction
  - 76,139 stay points
- Steps 2-4 – clustering and cluster refinement
  - ~6,500 places
  - Clustering metrics: Purity, entropy, NMI
- Step 5 – ranking
  - Ranking metrics: Precision@n, MAP, nDCG, Runtime

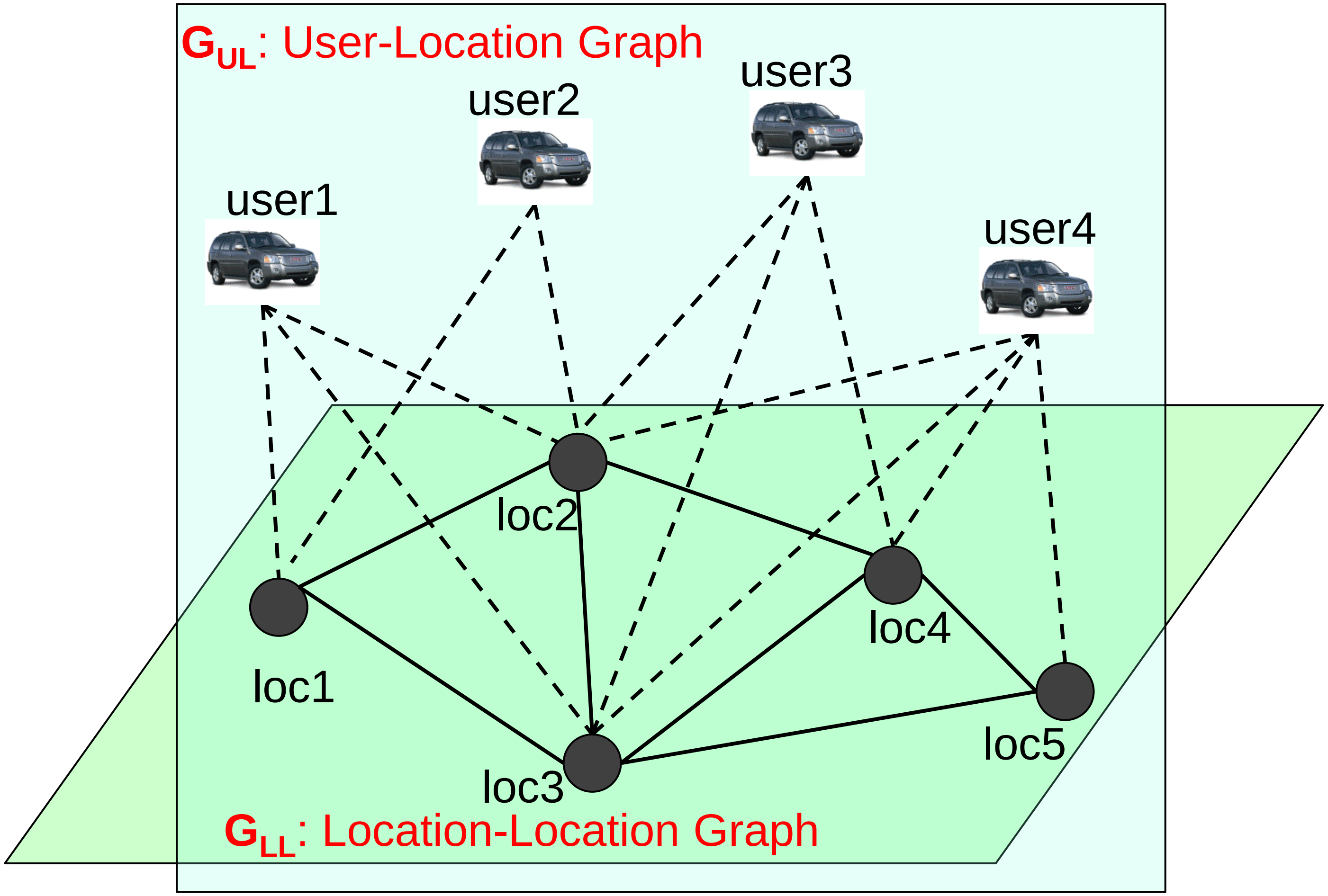
Ranking

- Exploit different aspects of the location records
  - The more visits, the more significant
  - The longer the durations of visits, the more significant
  - The more distinct visitors, the more significant
  - The longer the distances traveled to visit, the more significant

  - The more “near-by” significant places are, the more significant a place is.
  - The more a place is visited by objects that visit significant places, the more significant it is.

Two-Layered Graph

- **$\mathbf{G}_{LL}$** : a link represents a trip between two locations
- **$\mathbf{G}_{UL}$**: a link represents a visit of a user to a location


$\mathbf{G}_{\mathbf{UL}}$: User-Location Graph
user1
user2
user3
user4
loc1
loc2
loc3
loc4
loc5
$\mathbf{G}_{\mathbf{LL}}$: Location-Location Graph

Results

	Rank-by-visits	Rank-by-durations	HITS-based
MAP	0.2020	0.2126	0.062
P@20	0.45	0.45	0.1
P@50	0.36	0.38	0.12
nDCG@20	0.8261	0.8324	0.4555
nDCG@50	0.9678	0.7747	0.4380
Runtime (ms)	103	107	1536

	U-L	L-L	Unified	ST-Unified
MAP	0.3748	0.3020	0.4060	0.4274
P@20	0.75	0.6	0.9	0.95
P@50	0.68	0.52	0.74	0.76
nDCG@20	0.9411	0.9031	0.9678	0.9897
nDCG@50	0.9226	0.8827	0.9402	0.9717
Runtime (ms)	2209	3540	4234	4318

Directions Query Based Place Ranking

- How can we use directions queries for assigning significance to places and as a signal for the ranking of local search results?

- Directions query: $x \rightarrow y$ @ t
  - The user asks for directions from *x* to *y* at time *t*.

- Such queries will proliferate as navigation goes online.

- Idea: query $x \rightarrow y$ @ t is a vote that *y* is an important place.

Directions Query Based Place Ranking

- Exploit different aspects of the queries
- Count-based: The more queries to *y* @ *t*, the more significant *y* is (@ *t*).
- Distance-based: The longer the distances *x* →*y*, the more the more significant y is.
- Locality-based: The more queries *x* →*y*, the more significant *y* is for users close to *x*.

Experimental Study

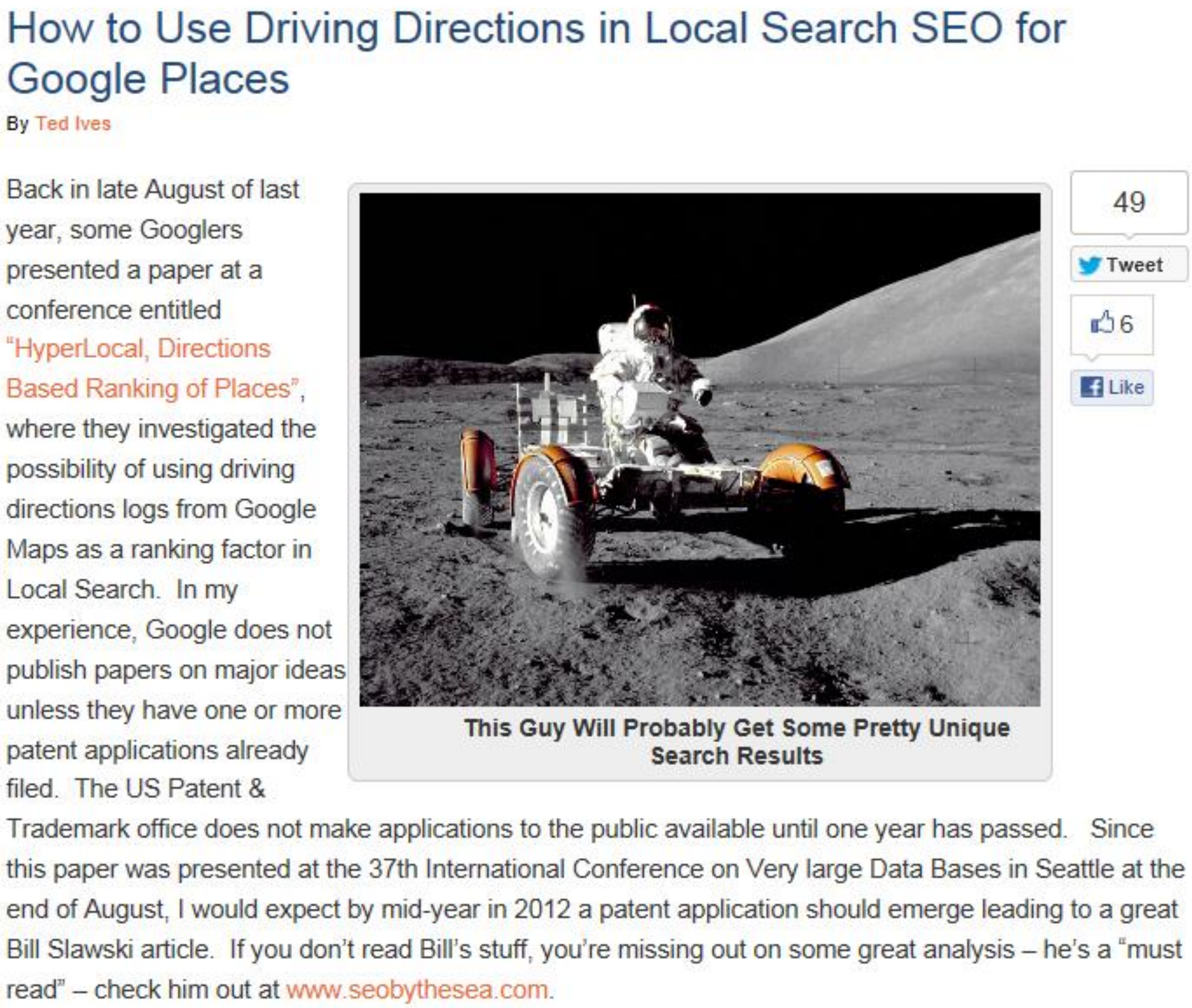
- Using query logs from Google

- The most obvious competitor is reviews and ratings.

- Similar quality as reviews
- Better coverage than reviews
- Better temporal granularity than reviews
  - Examples of finer temporal granularity: after-work bar, weekday lunch restaurant
  - Ability to better identify sentiment change

- A way of contending with review spam
  - Few queries and many positive reviews may signal spam

SEO Attention



How to Use Driving Directions in Local Search SEO for Google Places

By Ted Ives

Back in late August of last year, some Googlers presented a paper at a conference entitled "HyperLocal, Directions Based Ranking of Places", where they investigated the possibility of using driving directions logs from Google Maps as a ranking factor in Local Search. In my experience, Google does not publish papers on major ideas unless they have one or more patent applications already filed. The US Patent & Trademark office does not make applications to the public available until one year has passed. Since this paper was presented at the 37th International Conference on Very large Data Bases in Seattle at the end of August, I would expect by mid-year in 2012 a patent application should emerge leading to a great Bill Slawski article. If you don't read Bill's stuff, you're missing out on some great analysis – he's a "must read" – check him out at www.seobythesea.com.

This Guy Will Probably Get Some Pretty Unique Search Results

Blog post at www.coconutheadphones.com by Ted Ives

SEO for Best Practices

- Driving directions should
  - be from unique machines and unique users
  - be from a mix of mobile and desktop searches
  - be requested from different locations and distances
  - have a natural distribution of timing that match customer’s search patterns and the place’s opening hours
  - be from a mix of search entry paths (address search, product/service search)
- Searches from the location of the business are probably not helpful.
- If you obtain a lot of reviews without a lot of direction searches, that could be flagged as review spam.
- Don’t make directions too easy for your users.
  - Do not embed a form or a link on your website that generates a driving directions query. Any approach like this will probably be filtered out. In fact, if you provide such an experience, you’re actually hurting your rankings.

Based on a blog post at www.coconutheadphones.com by Ted Ives

Summary and Challenges

Summary

- The web is going mobile and has a spatial dimension.
- Many queries have local intent
- Spatial keyword queries
  - *k* nearest neighbor queries
  - Continuous *k* nearest neighbor queries
  - Using nearby relevant content for place ranking
  - Retrieve a set of objects that collectively best satisfy a query
- Use of UGC for place ranking
  - GPS records, directions queries

Challenges

- Structured queries and Amazon-style and social queries
  - Ample opportunities for much more customization of results
- Build in feedback mechanisms
  - "Figuring out how to build databases that get better the more people use them is actually the secret source of every Web 2.0 company" –Tim O'Reilly
- Tractability versus utility
  - The area is prone to NP completeness
- Avoid parameter overload
  - Problem vs. solution parameters
  - Hard-to-set, impossible-to-set parameters – relevance decreases exponentially with the number of such parameters
- User evaluation
  - Challenging – particularly for someone who used to study joins.

Acknowledgments and Readings

- Cao, X., L. Chen, G. Cong, C. S. Jensen, Q. Qu, A. Skovsgaard, D. Wu, and M. L. Yiu: Spatial Keyword Querying. ER, pp. 16-29 (2012)
- Wu, D., M. L. Yiu, G. Cong, and C. S. Jensen: Joint Top-K Spatial Keyword Query Processing. TKDE, to appear
- Cao, X., G. Cong, C. S. Jensen, J. J. Ng, B. C. Ooi, N.-T. Phan, D. Wu: SWORS: A System for the Efficient Retrieval of Relevant Spatial Web Objects. PVLDB, 5(12): 1914-1917 (2012)
- Wu, D., G. Cong, and C. S. Jensen: A Framework for Efficient Spatial Web Object Retrieval. VLDBJ, 26 pages, in online first
- Wu, D., M. L. Yiu, C. S. Jensen, G. Cong: Efficient Continuously Moving Top-K Spatial Keyword Query Processing. ICDE, pp. 541-552 (2011)
- Venetis, P., H. Gonzales, C. S. Jensen, A. Halevy: Hyper-Local, Directions-Based Ranking of Places. PVLDB 4(5): 290-301 (2011)
- Cao, X., G. Cong, C. S. Jensen, B. C. Ooi: Collective Spatial Keyword Querying. SIGMOD, pp. 373-384 (2011)
- Cao, X., G. Cong, C. S. Jensen: Retrieving Top-k Prestige-Based Relevant Spatial Web Objects. PVLDB 3(1): 373-384 (2010)
- Cao, X., G. Cong, C. S. Jensen: Mining Significant Semantic Locations From GPS Data. PVLDB 3(1): 1009-1020 (2010)
- Cong, G., C. S. Jensen, D. Wu: Efficient Retrieval of the Top-k Most Relevant Spatial Web Objects. PVLDB 2(1): 337-348 (2009)