Query Session Detection as a Cascade

Matthias Hagen Benno Stein Tino Rüb

Bauhaus-Universität Weimar
matthias.hagen@uni-weimar.de

SIR 2011
Dublin, Ireland
April 18, 2011

It's quiz time!

It's quiz time!

What is the user searching?

`paris hilton`

Without context . . .

`paris hilton`

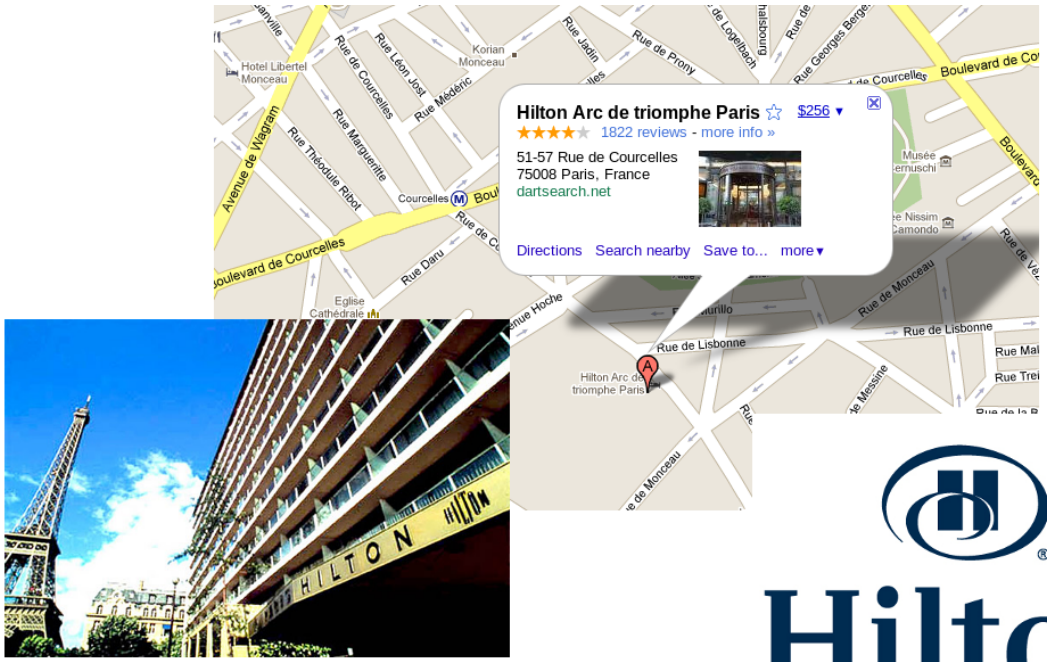
source: [http://upload.wikimedia.org/wikipedia/commons/2/26/Paris_Hilton_3_Crop.jpg]

What if you knew the previous queries?

```
paris hotels
paris marriott
paris hyatt
paris hilton
```

What if you knew the previous queries?

```
paris hotels
paris marriott
paris hyatt
paris hilton
```



sources: [http://www.alison-anderson.com/wp-content/uploads/hilton_hotel_paris_2.jpg]
[http://maps.google.com/]
[http://upload.wikimedia.org/wikipedia/en/e/eb/HI_mk_logo_hiltonbrandlogo.jpg]

Hilton
HOTELS & RESORTS

Query sessions: same information need

The benefits

- Improved understanding of user intent
- Improved retrieval performance via session knowledge

Query sessions: same information need

The benefits

- Improved understanding of user intent
- Improved retrieval performance via session knowledge

The "minor" issue

Users do not announce when querying for a new information need.

A typical query log

User	Query	Click domain + Click rank		Time
773	istanbul	en.wikipedia.org	1	2011-04-16 20:34:17
773	istanbul archeology			2011-04-17 12:02:54
773	istanbul archeology	www.kulturturizm.tr	6	2011-04-17 12:03:15
773	istanbul archeology	www.arkeoloji.gov.tr	13	2011-04-17 18:24:07
773	constantinople			2011-04-17 19:00:40
773	constantinople	www.roman-empire.net	4	2011-04-17 19:01:02
773	hurling			2011-04-17 19:03:01
773	hurling	en.wikipedia.org	1	2011-04-17 19:03:05
773	liam mccarthy cup			2011-04-17 23:33:04
773	liam mccarthy cup	www.hurling.net	5	2011-04-17 23:33:12
773	liam mccarthy cup	starbets.ie	16	2011-04-18 12:42:48

How to determine the break points?

User	Query	Click domain + Click rank		Time
773	istanbul	en.wikipedia.org	1	2011-04-16 20:34:17
773	istanbul archeology			2011-04-17 12:02:54
773	istanbul archeology	www.kulturturizm.tr	6	2011-04-17 12:03:15
773	istanbul archeology	www.arkeoloji.gov.tr	13	2011-04-17 18:24:07
773	constantinople			2011-04-17 19:00:40
773	constantinople	www.roman-empire.net	4	2011-04-17 19:01:02
773	hurling			2011-04-17 19:03:01
773	hurling	en.wikipedia.org	1	2011-04-17 19:03:05
773	liam mccarthy cup			2011-04-17 23:33:04
773	liam mccarthy cup	www.hurling.net	5	2011-04-17 23:33:12
773	liam mccarthy cup	starbets.ie	16	2011-04-18 12:42:48

The key is . . .

Automatic query session detection

Automatic query session detection

Usual "technique"

Check for consecutive queries whether same/new information need.

Example

773	istanbul	2011-04-16 20:34:17	✓ same
773	istanbul archeology	2011-04-17 18:24:07	✓ same
773	constantinople	2011-04-17 19:01:02	
	— — — — — — — — —		↯ new
773	hurling	2011-04-17 19:03:05	

Typical features

Temporal thresholds	5 minutes	[Silverstein et al., 1999]
	10–15 minutes	[He and Göker, 2000]
	30 minutes	[Downey et al., 2007]
	user specific	[Murray et al., 2006]
Lexical similarity	*n*-gram overlap	[Zhang and Moffat, 2006]
	Levenshtein distance	[Jones and Klinkner, 2008]
Semantic similarity	Search results	[Radlinski and Joachims, 2005]
	ESA	[Lucchese et al., 2011]

Previous methods

Observations

- Temporal thresholds: fast but bad accuracy
- Feature combinations: more accurate
- One of the best: Geometric method (time + lexical) [Gayo-Avello, 2009]

Previous methods

Observations

- Temporal thresholds: fast but bad accuracy
- Feature combinations: more accurate
- One of the best: Geometric method (time + lexical) [Gayo-Avello, 2009]

Shortcomings

- All features evaluated simultaneously → runtime
- Geometric method ignores semantics → accuracy

Examples

Subset test suffices

`hurling`
`hurling gaa` ✓ same

Geometric method fails

`hurling`
`mccarthy cup` ✓ same

We address the shortcomings in a cascade . . .

source: [http://wp.ltchambon.com/wp-content/uploads/2010/09/Cascade-de-Tufs-Baume-les-messieurs-Jura.jpg]

. . . well . . . a small 4-step cascade

source: [http://www.solarshop.com/solarpix/Solar Cascade 4 Tier GreenL.jpg]

. . . well . . . a small 4-step cascade

source: [http://www.solarshop.com/solarpix/Solar Cascade 4 Tier GreenL.jpg]

Step 1: Subset tests

↘

Step 2: Geometric method

↘

Step 3: ESA similarity

↙

Step 4: Search results

Basic Idea

Increased feature cost (runtime) from step to step.
Expensive features only if previous steps "unreliable."

Simple string comparison

Criterion

Consecutive queries q and q' in same session if q sub- or superset of q'.
Else: Goto Step 2.

Remarks: Repetition, specialization, or generalization.
Time gap = continuing a pending session.

Example

Repetition		Specialization		Generalization	
`hurling`	✓ same	`hurling`	✓ same	`hurling gaa`	✓ same
`hurling`		`hurling gaa`		`hurling`	

Combination of temporal and lexical features [Gayo-Avello, 2009]

For consecutive queries q and q'

f_{temp} = maximum of 0 and $1 - \frac{t}{24h}$ — t is time between q and q'

f_{lex} = cosine similarity of 3- to 5-grams of q' and s — s is session of q

Combination of temporal and lexical features [Gayo-Avello, 2009]

For consecutive queries q and q'

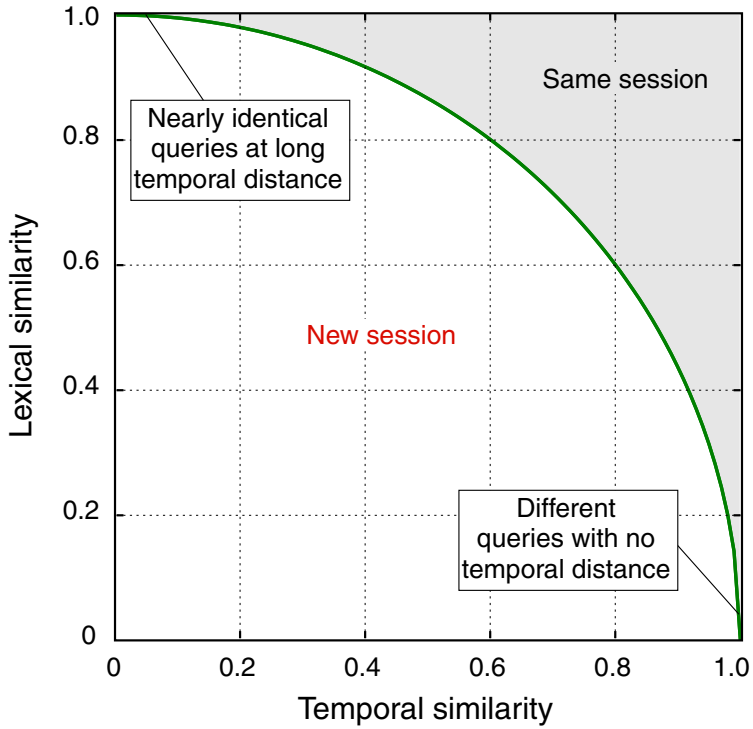
f_{temp} = maximum of 0 and $1 - \frac{t}{24h}$ — t is time between q and q'

f_{lex} = cosine similarity of 3- to 5-grams of q' and s — s is session of q

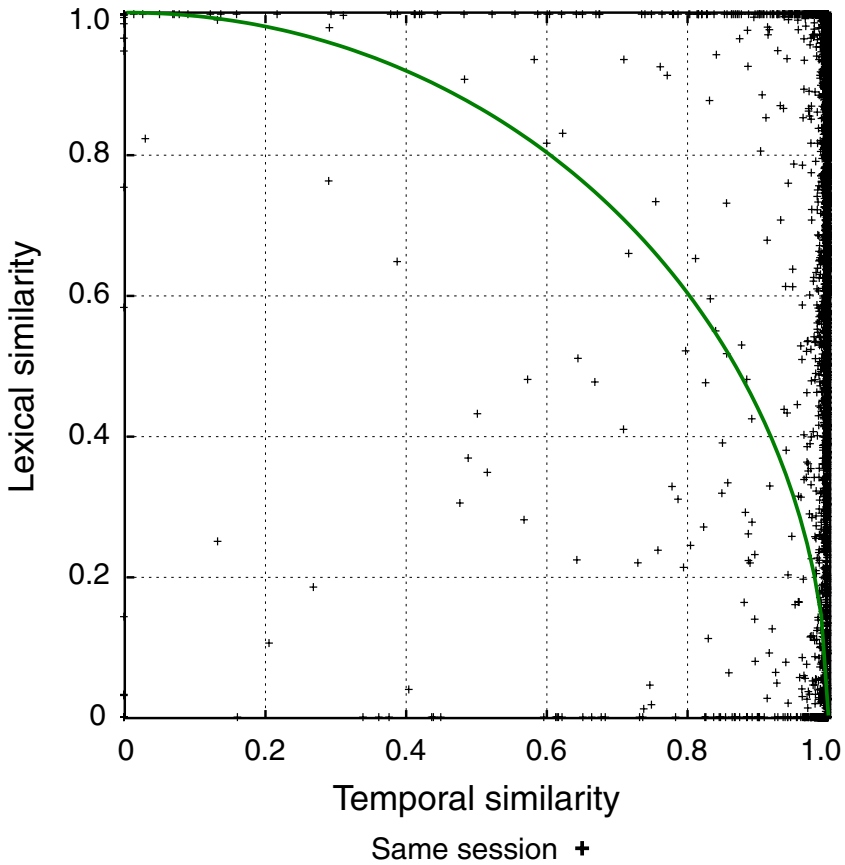
Criterion (original)

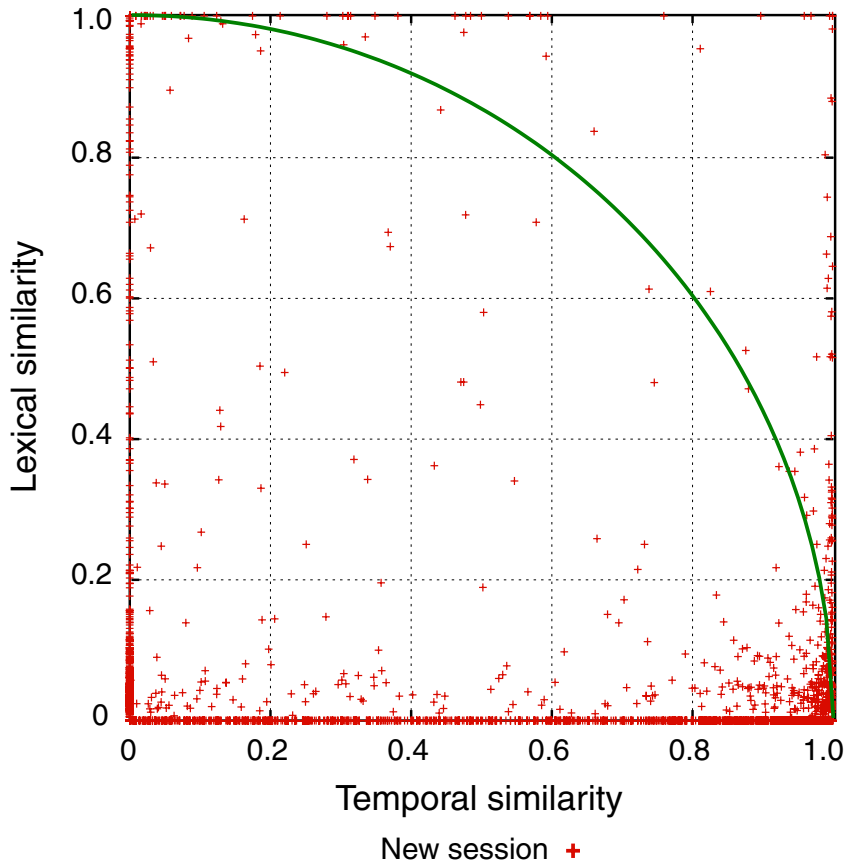
Consecutive queries q and q' in same session if

$$\sqrt{f_{\text{temp}}^2 + f_{\text{lex}}^2} \geq 1.$$

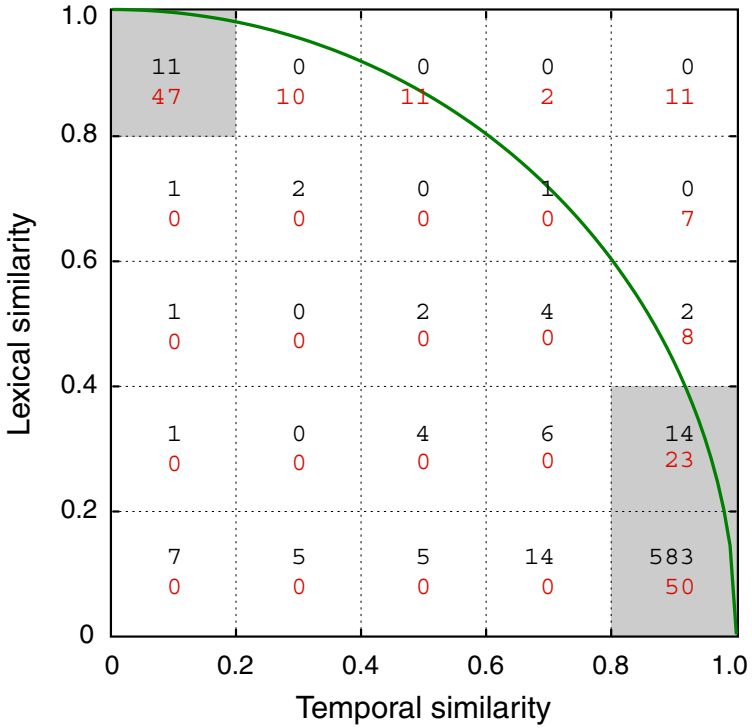


Performs well on standard test corpus . . .





. . . but has some problems "on the edge"



Major problems

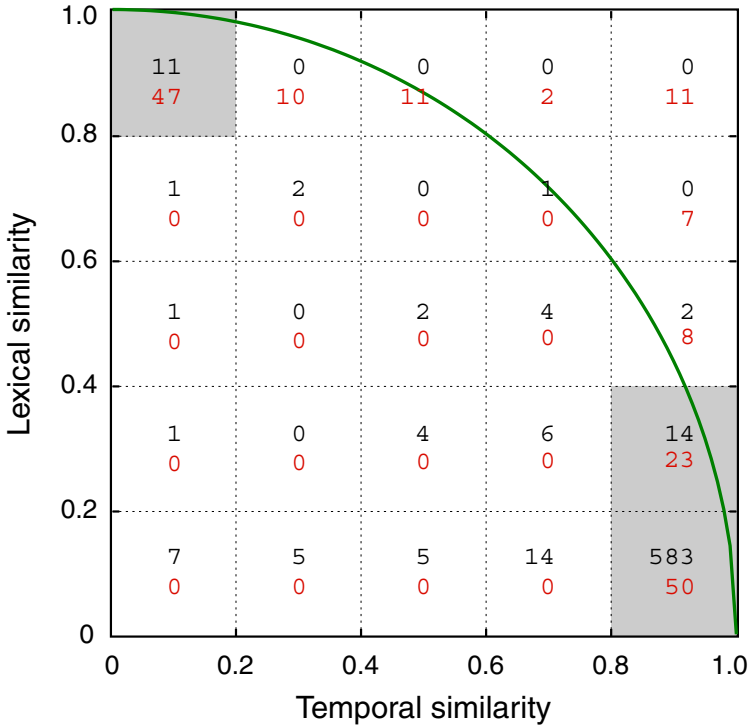
Similar queries, time gap (upper left)

$\rightarrow$ Merely a matter of opinion

Diff. queries, same semantics (lower right)

$\rightarrow$ Incorporate semantics

. . . but has some problems "on the edge"



Major problems

Similar queries, time gap (upper left)
$\rightarrow$ Merely a matter of opinion

Diff. queries, same semantics (lower right)
$\rightarrow$ Incorporate semantics

Criterion (adapted)

Original geometric method if $f_{\text{temp}} < 0.8$ or $f_{\text{lex}} > 0.4$.
Else: Goto Step 3.

How ESA works

[Gabrilovich and Markovitch, 2007]

Preprocessing

$tf \cdot idf$-weighted inverted index of Wikipedia articles $\rightarrow$ term-document matrix M

For consecutive queries q and q'

f_{esa} = cosine similarity of $M^T \cdot \mathbf{q}'$ and $M^T \cdot \mathbf{s}$ s is session of q

Criterion

Consecutive queries q and q' in same session if $f_{\text{esa}} \geq 0.35$.
Else: Goto Step 4.

Even more “semantics”

Idea

Enrich the short query strings with the results of some web search engine.

Criterion

Consecutive queries q and q' in same session iff they share at least one of the top 10 search results.

Even more "semantics"

Idea

Enrich the short query strings with the results of some web search engine.

Criterion

Consecutive queries q and q' in same session iff they share at least one of the top 10 search results.

Remark

If q and q' share no top 10 result, decision should be "not sure."

That's the complete cascade

source: [http://www.solarshop.com/solarpix/Solar Cascade 4 Tier GreenL.jpg]

Step 1: Subset tests

↘

Step 2: Geometric method

↘

Step 3: ESA similarity

↙

Step 4: Search results

That's the complete cascade

source: [http://www.solarshop.com/solarpix/Solar Cascade 4 Tier GreenL.jpg]

Step 1: Subset tests

↘

Step 2: Geometric method

↘

Step 3: ESA similarity

↙

Step 4: Search results

What about accuracy and performance?

Accuracy and runtime

Accuracy on Gayo-Avello's corpus (11 000 queries, 2.7 per session)

	Precision	Recall	F-Measure ($\beta = 1.5$)
Geometric	**0.8673**	0.9431	0.9184
Cascading	0.8618	**0.9676**	**0.9328**

Performance per step on Gayo-Avello's corpus

	affected	F-Measure	time	factor
Step 1	40.49%	0.8303	0.08 ms	1.0
Step 2	35.15%	0.9292	0.20 ms	2.5
Step 3	2.05%	0.9316	0.27 ms	3.4
Step 4	0.85%	0.9328	9.85 ms	123.1

Goal: high quality session test data

Our own use case

Sample sessions from the AOL log as test data.

AOL log (cleaned): 35.4 million interactions from 470 000 users.

Some figures

Step 4 involved on 22.5% $\rightarrow$ 8 million web queries

$\rightarrow$ 300 ms per search $\rightarrow$ 1 month

Goal: high quality session test data

Our own use case

Sample sessions from the AOL log as test data.
AOL log (cleaned): 35.4 million interactions from 470 000 users.

Some figures

Step 4 involved on 22.5% $\rightarrow$ 8 million web queries
$\rightarrow$ 300 ms per search $\rightarrow$ 1 month

Way out

- Drop Step 4 and the sessions on which it would have been invoked

Remaining sessions:	Cleaned AOL log:
F-Measure = 0.9755	27 minutes

Almost the end: The take-away messages!

What we have done

Results

- Cascading method
- Cheap features first
- Beats geometric
- 3 step version: simple, fast, high quality sessions

Future Work

- Postprocessing for multi-tasking
- Postprocessing for goals/missions

What we have (not) done

Results

- Cascading method
- Cheap features first
- Beats geometric
- 3 step version: simple, fast, high quality sessions

Future Work

- Postprocessing for multi-tasking
- Postprocessing for goals/missions

What we have (not) done

Results

- Cascading method
- Cheap features first
- Beats geometric
- 3 step version: simple, fast, high quality sessions

Future Work

- Postprocessing for multi-tasking
- Postprocessing for goals/missions

Thank you

☺