Query Session Detection as a Cascade

Matthias Hagen   Benno Stein   Tino Rüb

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

CIKM 2011
Glasgow, Scotland
October 25, 2011

It's quiz time!

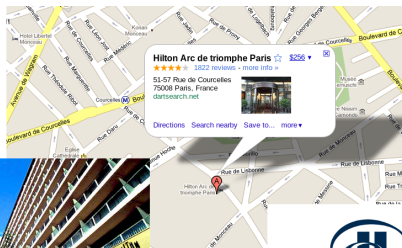What is the user searching?

`paris hilton`

paris hilton

What if you knew the previous queries?

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

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

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
42	`istanbul`	en.wikipedia.org	1	2011-10-22 20:34:17
42	`istanbul archeology`			2011-10-23 12:02:54
42	`istanbul archeology`	www.turizm.tr	6	2011-10-23 12:03:15
42	`istanbul archeology`	www.arkeoloji.tr	13	2011-10-23 18:24:07
42	`constantinople`			2011-10-23 19:12:40
42	`constantinople`	en.wikipedia.org	4	2011-10-23 19:13:02
42	`soccr glasgo`			2011-10-23 19:16:01
42	`soccer glasgow`			2011-10-23 19:16:11
42	`soccer glasgow`	www.soccer.uk	3	2011-10-23 19:16:15
42	`celtics vs rangers`			2011-10-23 20:33:04
42	`celtics vs rangers`	en.wikipedia.org	5	2011-10-23 20:33:12
42	`old firm`			2011-10-23 22:42:48

How to determine the break points?

User	Query	Click domain + Click rank		Time
42	istanbul	en.wikipedia.org	1	2011-10-22 20:34:17
42	istanbul archeology			2011-10-23 12:02:54
42	istanbul archeology	www.turizm.tr	6	2011-10-23 12:03:15
42	istanbul archeology	www.arkeoloji.tr	13	2011-10-23 18:24:07
42	constantinople			2011-10-23 19:12:40
42	constantinople	en.wikipedia.org	4	2011-10-23 19:13:02
		— — — — — — — — — — — — — — —		
42	soccr glasgo			2011-10-23 19:16:01
42	soccer glasgow			2011-10-23 19:16:11
42	soccer glasgow	www.soccer.uk	3	2011-10-23 19:16:15
42	celtics vs rangers			2011-10-23 20:33:04
42	celtics vs rangers	en.wikipedia.org	5	2011-10-23 20:33:12
42	old firm			2011-10-23 22:42:48

Automatic query session detection

Automatic query session detection

Usual "technique"

Check for consecutive queries whether same/new information need.

Example

42	`istanbul`	2011-10-22 20:34:17	✓ same
42	`istanbul archeology`	2011-10-23 18:24:07	✓ same
42	`constantinople`	2011-10-23 19:12:40	
	— — — — — — — — —		⚡ new
42	`soccer glasgow`	2011-10-23 19:16:11	

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]

Feature combinations

- More accurate than single features
- One of the best: Geometric method (time + lexical) [Gayo-Avello, 2009]

Previous methods

Feature combinations

- More accurate than single features
- One of the best: Geometric method (time + lexical)   [Gayo-Avello, 2009]

Shortcomings

- All features evaluated simultaneously $\rightarrow$ runtime
- Geometric method ignores semantics $\rightarrow$ accuracy

Examples

Subset test suffices

soccer
soccer glasgow   ✓ same

Geometric method fails

celtics vs rangers
old firm   ✓ same

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

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 test

     ↘

    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."

Step 1: Subset test

User	Query	Click domain + Click rank		Time
42	`istanbul`	en.wikipedia.org	1	2011-10-22 20:34:17
42	`istanbul archeology`			2011-10-23 12:02:54
42	`istanbul archeology`	www.turizm.tr	6	2011-10-23 12:03:15
42	`istanbul archeology`	www.arkeoloji.tr	13	2011-10-23 18:24:07
42	`constantinople`			2011-10-23 19:12:40
42	`constantinople`	en.wikipedia.org	4	2011-10-23 19:13:02
42	`soccr glasgo`			2011-10-23 19:16:01
42	`soccer glasgow`			2011-10-23 19:16:11
42	`soccer glasgow`	www.soccer.uk	3	2011-10-23 19:16:15
42	`celtics vs rangers`			2011-10-23 20:33:04
42	`celtics vs rangers`	en.wikipedia.org	5	2011-10-23 20:33:12
42	`old firm`			2011-10-23 22:42:48

User	Query	Click domain + Click rank		Time
42	`istanbul`	en.wikipedia.org	1	2011-10-22 20:34:17
42	`istanbul archeology`			2011-10-23 12:02:54
42	`istanbul archeology`	www.turizm.tr	6	2011-10-23 12:03:15
42	`istanbul archeology`	www.arkeoloji.tr	13	2011-10-23 18:24:07
— — — —	— — — — — — —	— — — — — — — — — — — —		— — —
42	`constantinople`			2011-10-23 19:12:40
42	`constantinople`	en.wikipedia.org	4	2011-10-23 19:13:02
— — — —	— — — — — — —	— — — — — — — — — — — —		— — —
42	`soccr glasgo`			2011-10-23 19:16:01
42	`soccer glasgow`			2011-10-23 19:16:11
42	`soccer glasgow`	www.soccer.uk	3	2011-10-23 19:16:15
— — — —	— — — — — — —	— — — — — — — — — — — —		— — —
42	`celtics vs rangers`			2011-10-23 20:33:04
42	`celtics vs rangers`	en.wikipedia.org	5	2011-10-23 20:33:12
— — — —	— — — — — — —	— — — — — — — — — — — —		— — —
42	`old firm`			2011-10-23 22:42:48

Step 3: Explicit Semantic Analysis [Gabrilovich and Markovitch, 2007]

User	Query	Click domain +	Click rank	Time
42	istanbul	en.wikipedia.org	1	2011-10-22 20:34:17
42	istanbul archeology			2011-10-23 12:02:54
42	istanbul archeology	www.turizm.tr	6	2011-10-23 12:03:15
42	istanbul archeology	www.arkeoloji.tr	13	2011-10-23 18:24:07
42	constantinople			2011-10-23 19:12:40
42	constantinople	en.wikipedia.org	4	2011-10-23 19:13:02
	— — — — — — — — — — — — — — — — — —			
42	soccr glasgo			2011-10-23 19:16:01
42	soccer glasgow			2011-10-23 19:16:11
42	soccer glasgow	www.soccer.uk	3	2011-10-23 19:16:15
42	celtics vs rangers			2011-10-23 20:33:04
42	celtics vs rangers	en.wikipedia.org	5	2011-10-23 20:33:12
	— — — — — — — — — — — — — — — — — —			
42	old firm			2011-10-23 22:42:48

Step 4: Search results

User	Query	Click domain + Click rank		Time
42	`istanbul`	en.wikipedia.org	1	2011-10-22 20:34:17
42	`istanbul archeology`			2011-10-23 12:02:54
42	`istanbul archeology`	www.turizm.tr	6	2011-10-23 12:03:15
42	`istanbul archeology`	www.arkeoloji.tr	13	2011-10-23 18:24:07
42	`constantinople`			2011-10-23 19:12:40
42	`constantinople`	en.wikipedia.org	4	2011-10-23 19:13:02
	— — — — — — — — — — — — — — — — — — —			
42	`soccr glasgo`			2011-10-23 19:16:01
42	`soccer glasgow`			2011-10-23 19:16:11
42	`soccer glasgow`	www.soccer.uk	3	2011-10-23 19:16:15
42	`celtics vs rangers`			2011-10-23 20:33:04
42	`celtics vs rangers`	en.wikipedia.org	5	2011-10-23 20:33:12
42	`old firm`			2011-10-23 22:42:48

That's the complete cascade

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

Step 1: Subset test

      Step 2: Geometric method

          Step 3: ESA similarity

Step 4: Search results

That's the complete cascade

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

Step 1: Subset test

      ↘

    Step 2: Geometric method

          ↘

          Step 3: ESA similarity

        ↙

Step 4: Search results

What about accuracy and runtime?

Experimental Evaluation

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

	decides	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

Experimental Evaluation

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

	decides	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

Remark: Without Step 4 about 2 700 queries per second!

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

☺