Improved Cascade for Search Mission Detection

Matthias Hagen Jakob Gomoll Benno Stein

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

SIR 2012
Barcelona, Spain
April 1, 2012

What is the user searching?

bar celona

Without context . . .

bar celona

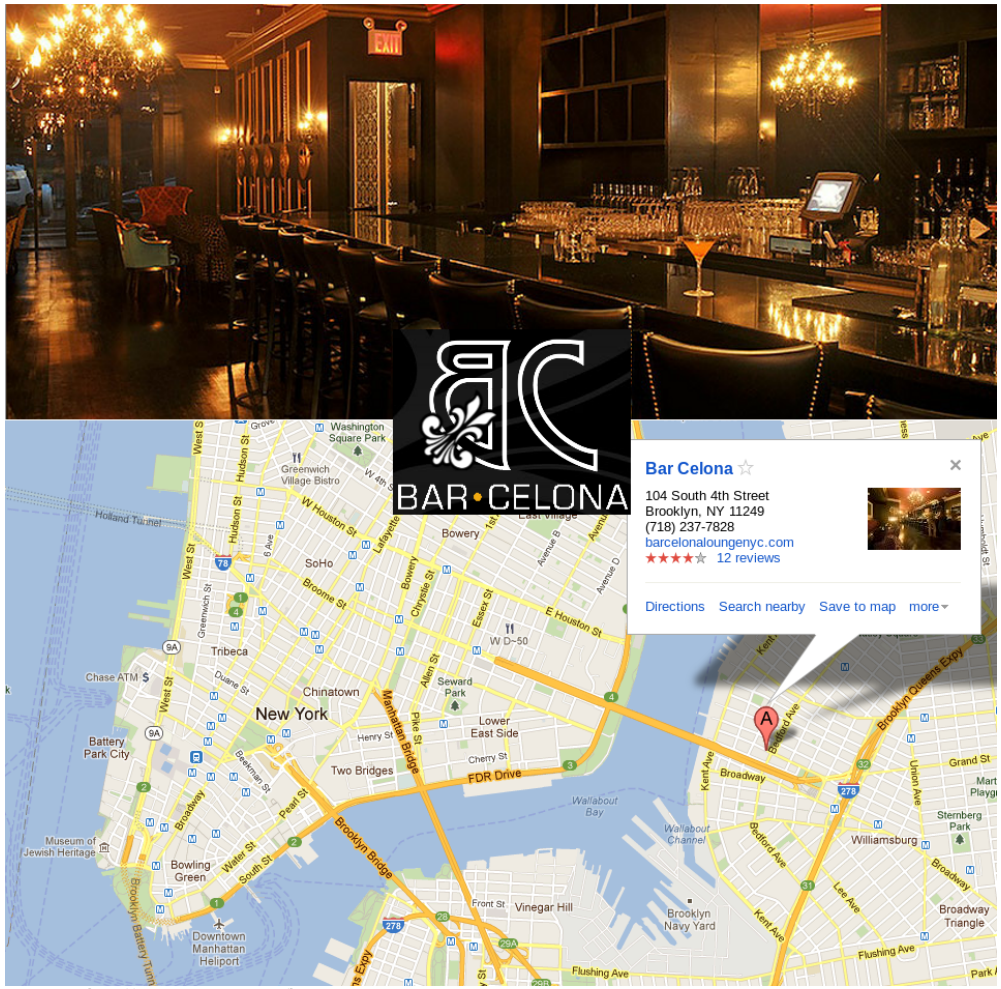
source: [http://ecir2012.upf.edu/images/header.jpg]

What if you knew the previous queries?

```
new york nightlife
new york clubs
new york bars
bar celona
```

What if you knew the previous queries?

```
new york nightlife
new york clubs
new york bars
bar celona
```



sources: [http://barcelonaloungenyc.com/] [http://maps.google.com]

Query sessions: same information need

Knowing sessions can improve

- Understanding of user intent
- Retrieval performance

A typical query log

User	Query	Click domain	Click rank	Time
42	istanbul	en.wikipedia.org	1	2012-03-22 20:34:17
42	istanbul archeology			2012-03-23 12:02:54
42	istanbul archeology	www.turizm.tr	6	2012-03-23 12:03:15
42	istanbul archeology	www.arkeoloji.tr	13	2012-03-23 18:24:07
42	constantinople			2012-03-23 19:12:40
42	constantinople	en.wikipedia.org	4	2012-03-23 19:13:02
42	football barclona			2012-03-23 19:16:01
42	football barcelona			2012-03-23 19:16:11
42	football barcelona	www.football.es	3	2012-03-23 19:16:15
42	real vs barca			2012-03-23 20:33:04
42	real vs barca	en.wikipedia.org	5	2012-03-23 20:33:12
42	el clasico			2012-03-23 22:42:48
42	constantinople			2012-03-24 10:17:09

Highlighted sessions

User	Query	Click domain + Click rank		Time
42	istanbul	en.wikipedia.org	1	2012-03-22 20:34:17
42	istanbul archeology			2012-03-23 12:02:54
42	istanbul archeology	www.turizm.tr	6	2012-03-23 12:03:15
42	istanbul archeology	www.arkeoloji.tr	13	2012-03-23 18:24:07
42	constantinople			2012-03-23 19:12:40
42	constantinople	en.wikipedia.org	4	2012-03-23 19:13:02
42	football barclona			2012-03-23 19:16:01
42	football barcelona			2012-03-23 19:16:11
42	football barcelona	www.football.es	3	2012-03-23 19:16:15
42	real vs barca			2012-03-23 20:33:04
42	real vs barca	en.wikipedia.org	5	2012-03-23 20:33:12
42	el clasico			2012-03-23 22:42:48
42	constantinople			2012-03-24 10:17:09

Multitasking and search missions

Observations [Spink et al., 2006; Jones and Klinkner, 2008]

- Search intents interleaved — Multitasking
- Long-term tasks with several sessions — Search missions

Multitasking and search missions

Observations [Spink et al., 2006; Jones and Klinkner, 2008]

- Search intents interleaved — Multitasking
- Long-term tasks with several sessions — Search missions

Session detection

Focused on consecutive queries → Misses multitasking/missions

Example

```
42  istanbul                2012-03-22 20:34:17   same ✓
42  istanbul archeology     2012-03-23 18:24:07
    — — — — — — — — — —                           new  ✓
42  football barcelona      2012-03-23 19:16:11
    — — — — — — — — — —                           new  ↯
42  constantinople          2012-03-24 10:17:09
```

Our topic . . .

Session detection + Multitasking/missions

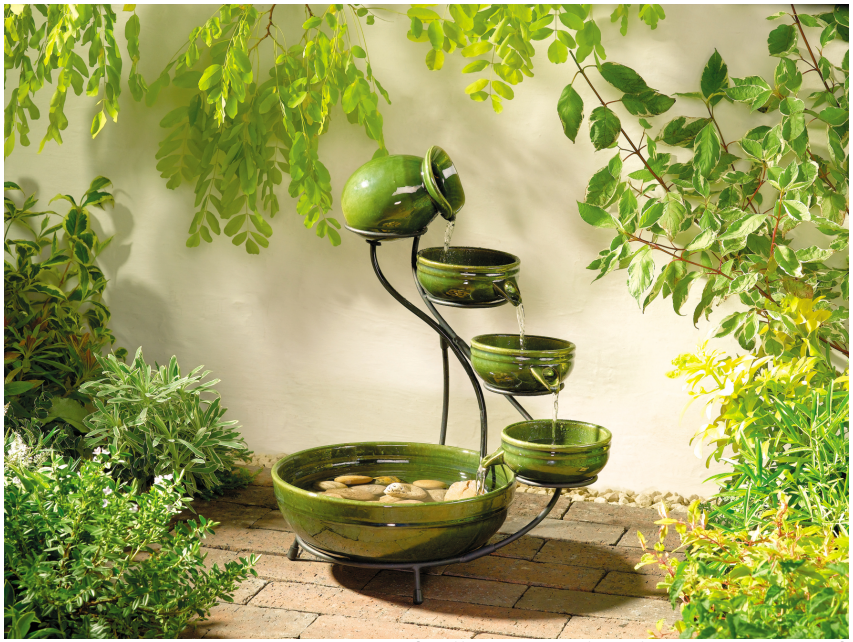
Typical query similarity features

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]
	Linked Open Data	[Hollink et al., 2011]

Our last year's cascade . . . [Hagen et al., 2011]

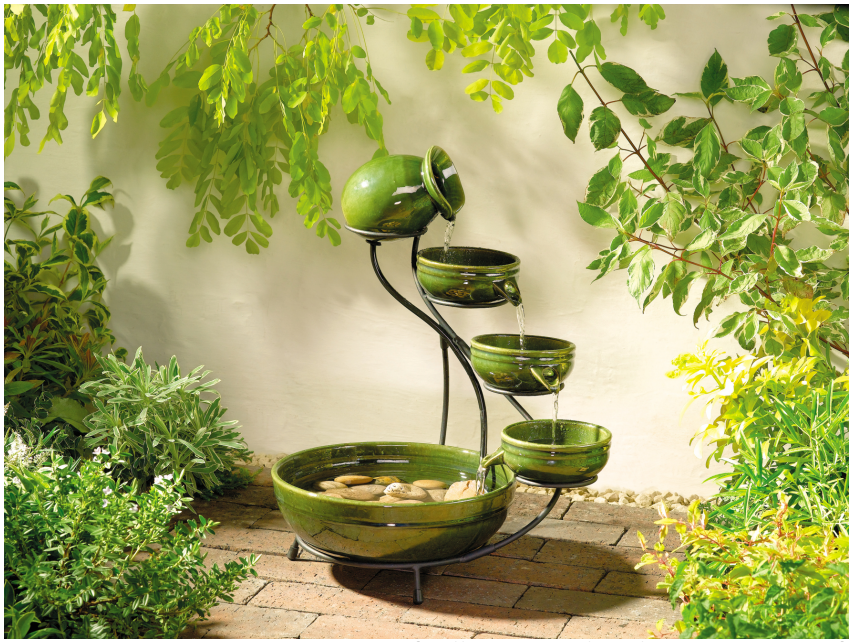
source: [http://wp.ltchambon.com/wp-content/uploads/2010/09/Cascade-de-Tufs-Baume-les-messieurs-Jura.jpg]

. . . well . . . it looks more like this [Hagen et al., 2011]



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

. . . well . . . it looks more like this [Hagen et al., 2011]



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

Step 1: Subset test

↘

Step 2: Geometric method

↘

Step 3: ESA similarity

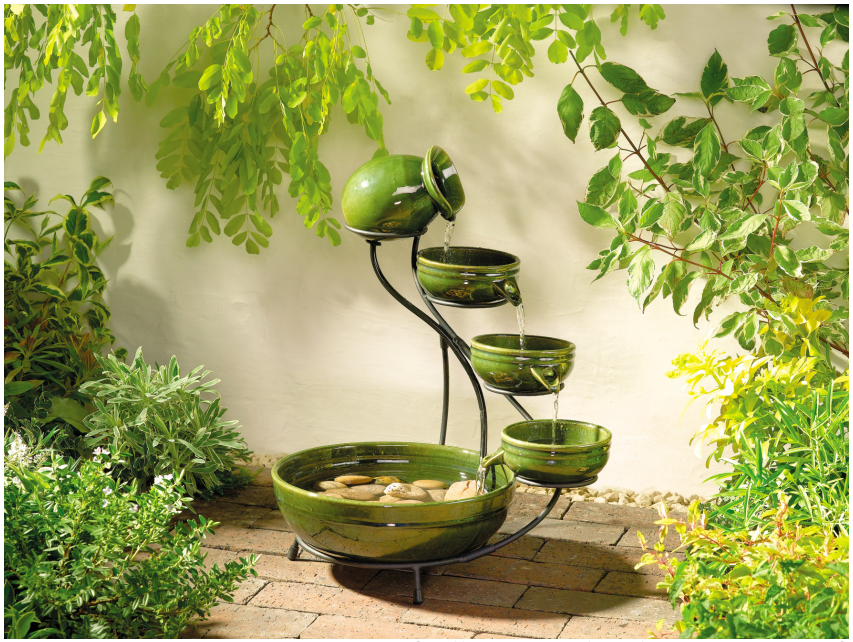
↙

Step 4: Search Results

Basic Idea

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

. . . well . . . it looks more like this (improved)



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: Linked Open Data

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	2012-03-22 20:34:17
42	istanbul archeology			2012-03-23 12:02:54
42	istanbul archeology	www.turizm.tr	6	2012-03-23 12:03:15
42	istanbul archeology	www.arkeoloji.tr	13	2012-03-23 18:24:07
42	constantinople			2012-03-23 19:12:40
42	constantinople	en.wikipedia.org	4	2012-03-23 19:13:02
42	football barclona			2012-03-23 19:16:01
42	football barcelona			2012-03-23 19:16:11
42	football barcelona	www.football.es	3	2012-03-23 19:16:15
42	real vs barca			2012-03-23 20:33:04
42	real vs barca	en.wikipedia.org	5	2012-03-23 20:33:12
42	el clasico			2012-03-23 22:42:48
42	constantinople			2012-03-24 10:17:09

Step 2: Geometric method

[Gayo-Avello, 2009]

User	Query	Click domain + Click rank		Time
42	istanbul	en.wikipedia.org	1	2012-03-22 20:34:17
42	istanbul archeology			2012-03-23 12:02:54
42	istanbul archeology	www.turizm.tr	6	2012-03-23 12:03:15
42	istanbul archeology	www.arkeoloji.tr	13	2012-03-23 18:24:07
42	constantinople			2012-03-23 19:12:40
42	constantinople	en.wikipedia.org	4	2012-03-23 19:13:02
42	football barclona			2012-03-23 19:16:01
42	football barcelona			2012-03-23 19:16:11
42	football barcelona	www.football.es	3	2012-03-23 19:16:15
42	real vs barca			2012-03-23 20:33:04
42	real vs barca	en.wikipedia.org	5	2012-03-23 20:33:12
42	el clasico			2012-03-23 22:42:48
42	constantinople			2012-03-24 10:17:09

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

User	Query	Click domain + Click rank		Time
42	istanbul	en.wikipedia.org	1	2012-03-22 20:34:17
42	istanbul archeology			2012-03-23 12:02:54
42	istanbul archeology	www.turizm.tr	6	2012-03-23 12:03:15
42	istanbul archeology	www.arkeoloji.tr	13	2012-03-23 18:24:07
42	constantinople			2012-03-23 19:12:40
42	constantinople	en.wikipedia.org	4	2012-03-23 19:13:02
42	football barclona			2012-03-23 19:16:01
42	football barcelona			2012-03-23 19:16:11
42	football barcelona	www.football.es	3	2012-03-23 19:16:15
42	real vs barca			2012-03-23 20:33:04
42	real vs barca	en.wikipedia.org	5	2012-03-23 20:33:12
42	el clasico			2012-03-23 22:42:48
42	constantinople			2012-03-24 10:17:09

Step 4: Linked Open Data connections [Hollink et al., 2011]

User	Query	Click domain + Click rank		Time
42	istanbul	en.wikipedia.org	1	2012-03-22 20:34:17
42	istanbul archeology			2012-03-23 12:02:54
42	istanbul archeology	www.turizm.tr	6	2012-03-23 12:03:15
42	istanbul archeology	www.arkeoloji.tr	13	2012-03-23 18:24:07
42	constantinople			2012-03-23 19:12:40
42	constantinople	en.wikipedia.org	4	2012-03-23 19:13:02
42	football barclona			2012-03-23 19:16:01
42	football barcelona			2012-03-23 19:16:11
42	football barcelona	www.football.es	3	2012-03-23 19:16:15
42	real vs barca			2012-03-23 20:33:04
42	real vs barca	en.wikipedia.org	5	2012-03-23 20:33:12
42	el clasico			2012-03-23 22:42:48
42	constantinople			2012-03-24 10:17:09

What about multitasking/missions?

Idea

Run the cascade twice:

1. Session detection on query level
2. Multitasking/mission detection on session level

First run: detected sessions

User	Query	Click domain + Click rank		Time
42	istanbul	en.wikipedia.org	1	2012-03-22 20:34:17
42	istanbul archeology			2012-03-23 12:02:54
42	istanbul archeology	www.turizm.tr	6	2012-03-23 12:03:15
42	istanbul archeology	www.arkeoloji.tr	13	2012-03-23 18:24:07
42	constantinople			2012-03-23 19:12:40
42	constantinople	en.wikipedia.org	4	2012-03-23 19:13:02
42	football barclona			2012-03-23 19:16:01
42	football barcelona			2012-03-23 19:16:11
42	football barcelona	www.football.es	3	2012-03-23 19:16:15
42	real vs barca			2012-03-23 20:33:04
42	real vs barca	en.wikipedia.org	5	2012-03-23 20:33:12
42	el clasico			2012-03-23 22:42:48
42	constantinople			2012-03-24 10:17:09

Second run: multitasking/mission detection

User	Query	Click domain + Click rank		Time
42	istanbul	en.wikipedia.org	1	2012-03-22 20:34:17
42	istanbul archeology			2012-03-23 12:02:54
42	istanbul archeology	www.turizm.tr	6	2012-03-23 12:03:15
42	istanbul archeology	www.arkeoloji.tr	13	2012-03-23 18:24:07
42	constantinople			2012-03-23 19:12:40
42	constantinople	en.wikipedia.org	4	2012-03-23 19:13:02
42	football barclona			2012-03-23 19:16:01
42	football barcelona			2012-03-23 19:16:11
42	football barcelona	www.football.es	3	2012-03-23 19:16:15
42	real vs barca			2012-03-23 20:33:04
42	real vs barca	en.wikipedia.org	5	2012-03-23 20:33:12
42	el clasico			2012-03-23 22:42:48
42	constantinople			2012-03-24 10:17:09

Second run: multitasking/mission detection

User	Query	Click domain + Click rank		Time
42	istanbul	en.wikipedia.org	1	2012-03-22 20:34:17
42	istanbul archeology			2012-03-23 12:02:54
42	istanbul archeology	www.turizm.tr	6	2012-03-23 12:03:15
42	istanbul archeology	www.arkeoloji.tr	13	2012-03-23 18:24:07
42	constantinople			2012-03-23 19:12:40
42	constantinople	en.wikipedia.org	4	2012-03-23 19:13:02
42	football barclona			2012-03-23 19:16:01
42	football barcelona			2012-03-23 19:16:11
42	football barcelona	www.football.es	3	2012-03-23 19:16:15
42	real vs barca			2012-03-23 20:33:04
42	real vs barca	en.wikipedia.org	5	2012-03-23 20:33:12
42	el clasico			2012-03-23 22:42:48
42	constantinople			2012-03-24 10:17:09

What about accuracy and runtime?

Available evaluation corpora

Gayo-Avello's session detection corpus (AOL log, 1 annotator)

- 11 500 queries — But: empty queries, order changed
- 215 users — But: many with ≤ 3 queries
- 2.7 queries per session — But: several annotation errors

Lucchese et al.'s mission detection corpus (AOL log, 1 annotator)

- 1500 queries — But: 97% of queries dropped
- 13 users

Our new mission detection corpus (basis: Gayo-Avello, 2 annotators)

- 8800 queries — Empty/URL queries removed
- 127 users — Users with ≤ 3 queries removed
- 11 missions per user with 6.33 queries

Available evaluation corpora

Gayo-Avello's session detection corpus (AOL log, 1 annotator)

- 11 500 queries — But: empty queries, order changed
- 215 users — But: many with ≤ 3 queries
- 2.7 queries per session — But: several annotation errors

Lucchese et al.'s mission detection corpus (AOL log, 1 annotator)

- 1500 queries — But: 97% of queries dropped
- 13 users

Our new mission detection corpus (basis: Gayo-Avello, 2 annotators)

- 8800 queries — Empty/URL queries removed
- 127 users — Users with ≤ 3 queries removed
- 11 missions per user with 6.33 queries

Available evaluation corpora

Gayo-Avello's session detection corpus (AOL log, 1 annotator)

- 11 500 queries — But: empty queries, order changed
- 215 users — But: many with ≤ 3 queries
- 2.7 queries per session — But: several annotation errors

Lucchese et al.'s mission detection corpus (AOL log, 1 annotator)

- 1500 queries — But: 97% of queries dropped
- 13 users

Our new mission detection corpus (basis: Gayo-Avello, 2 annotators)

- 8800 queries — Empty/URL queries removed
- 127 users — Users with ≤ 3 queries removed
- 11 missions per user with 6.33 queries

Accuracy and runtime

Session accuracy on our corpus (6630 queries, 25 % training)

	F-Measure	Runtime
Original cascade (3 steps)	0.875	100 %
Improved cascade (3 steps)	**0.890**	90 %
Improved cascade (4 steps)	**0.890**	$\gg$100 %

Mission accuracy on our corpus (6630 queries, 25 % training)

- 556 continuations correctly detected (170 missed)
- 97 sessions wrongly assigned a continuation F-Measure 0.798

Accuracy and runtime

Session accuracy on our corpus (6630 queries, 25 % training)

	F-Measure	Runtime
Original cascade (3 steps)	0.875	100 %
Improved cascade (3 steps)	**0.890**	90 %
Improved cascade (4 steps)	**0.890**	$\gg$100 %

Mission accuracy on our corpus (6630 queries, 25 % training)

- 556 continuations correctly detected (170 missed)
- 97 sessions wrongly assigned a continuation F-Measure 0.798

Observations

- Cascade applicable to mission detection
- Linked Open Data not that useful yet

Almost the end: The take-away messages!

What we have done

Results

- Improved cascading method
- Cheap features first
- Applicable to mission detection
- LOD not really useful yet
- Large mission corpus

Future Work

- Prune LOD graph
- Index complete Wikipedia
- WordNet

What we have (not) done

Results

- Improved cascading method
- Cheap features first
- Applicable to mission detection
- LOD not really useful yet
- Large mission corpus

Future Work

- Prune LOD graph
- Index complete Wikipedia
- WordNet

What we have (not) done

Results

- Improved cascading method
- Cheap features first
- Applicable to mission detection
- LOD not really useful yet
- Large mission corpus

Future Work

- Prune LOD graph
- Index complete Wikipedia
- WordNet

Thank you

☺