Query segmentation revisited

Matthias Hagen Martin Potthast Benno Stein Christof Bräutigam

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

WWW 2011
Hyderabad, India
March 31, 2011

It's quiz time!

It's quiz time!

What is the user searching?

`new york times square dance`

Is it: new york times square dance ?

Image source: [http://www.theepochtimes.com/n2/images/stories/large/2009/08/06/Bollywood1.jpg]

Is it: new york times square dance ?!

"All the News That's Fit to Print"

The New York Times

Late Edition
Today, a shower, yielding to sun, high 64. Tonight, chilly, low 45. Tomorrow, sunny and cool, high 64. Yesterday's high, 70, low, 54. Weather map and details, Page 24.

VOL. CLVIII . No. 54,678 © 2009 The New York Times NEW YORK, SUNDAY, MAY 17, 2009 $5 beyond the greater New York metropolitan area. $4.00

Square Dance

19 U.S. States Have Designated It As Their Official State Dance

From a Theory To a Consensus On Emissions

Permits Gain Political Edge Over Taxation

By JOHN M. BRODER

WASHINGTON — As Congress weighs imposing a mandatory limit on climate-altering gases — an outcome still far from certain — it is likely to turn to a system that sets a government ceiling on total emissions and allows polluting industries to buy and sell permits to meet it.

That approach, known as cap and trade, has been embraced by President Obama, Democratic leaders in Congress, mainstream environmental groups and a growing number of business interests, including energy-consuming industries like autos, steel and aluminum.

But not long ago, many of today's supporters dismissed the idea of tradable emissions permits as an industry-inspired Republican scheme to avoid the real costs of cutting air pollution. The right answer, they said, was strict government regulation, state-of-the-art technology and a federal tax on every ton of harmful emissions.

How did cap and trade, hatched as an academic theory in obscure economic journals half a century ago, become the policy of choice in the debate over how to slow the heating of the planet? And how did it come to eclipse the idea of simply slapping a tax on energy consumption that befouls the public square or leaves the nation hostage to foreign oil

CONSERVATIVES MAP STRATEGIES ON COURT FIGHT

MEMOS OUTLINE ATTACKS

Hoping to Re-Energize G.O.P. by Opposing Obama's Choice

By CHARLIE SAVAGE

WASHINGTON — If President Obama nominates Judge Diane P. Wood to the Supreme Court, conservatives plan to attack her as an "outspoken" supporter of "abortion, including partial-birth abortion."

If he nominates Judge Sonia Sotomayor, they plan to accuse her of being "willing to expand constitutional rights beyond the text of the Constitution."

And if he nominates Kathleen M. Sullivan, a law professor at Stanford, they plan to denounce her as a "prominent supporter of homosexual marriage."

Preparing to oppose the confirmation of Mr. Obama's eventual choice to succeed Justice David H. Souter, who is retiring, conservative groups are working together to stockpile ammunition. Ten memorandums summarizing their research, obtained by The New York Times, provide a window onto how they hope to frame the coming debate.

The memorandums dissect possible nominees' records, noting statements the groups find objectionable on issues like abortion, same-sex marriage, the sep-

Image sources: [http://blog.caseytempleton.com/wp-content/uploads/2009/05/090517_nytfrontpage1.jpg] [http://upload.wikimedia.org/wikipedia/commons/0/03/Square_Dance_Group.jpg]

Segment your queries!

The benefits

- Improved retrieval precision
- Potential disambiguation
- Reformulations on segment level

The syntax

Quotes around segments: `"new york" "times square" dance`

Segment your queries!

The benefits

- Improved retrieval precision
- Potential disambiguation
- Reformulations on segment level

The syntax

Quotes around segments: `"new york" "times square" dance`

The "minor" issue ...

Most web searchers are not even aware of the quotes option

The way out . . .

Automatic pre-retrieval query segmentation

The way out . . .

Automatic pre-retrieval query segmentation

Remark: Runtime is crucial!

The computational problem as we see it

Query Segmentation

- Given a keyword query
- Find the "best" segmentation

Remarks: We assume correct spelling!
We do not change keywords!

Example

Given the query	`new york times square dance`
Solutions could be	`"new york" "times square" dance`
	`"new york times" "square dance"`
But not (word order!)	`"new york" "dance times square"`

(a Latin dance studio in NYC)

Standing on the shoulders of . . .

Mutual information	[Risvik et al., WWW 2003] [Jones et al., WWW 2006] [Huang et al., WWW 2010]
Supervised learning	[Bergsma and Wang, EMNLP-CoNLL 2007] [Bendersky et al., SIGIR 2009]
Unsupervised learning	[Tan and Peng, WWW 2008] [Zhang et al., ACL-IJCNLP 2009]
Retrieval feedback	[Brenes et al., CERI 2010] [Bendersky et al., CIKM 2010]
Query log	[Mishra et al., WWW 2011]
Naïve	[Hagen et al., SIGIR 2010]

KISS – Keep it simple and stupid!

Image source: [http://1.bp.blogspot.com/_UDZXrzYpS4k/THRrh8KPvVI/AAAAAAAAAoc/Be1HjlnRy1c/s400/lipstick-mirror.jpg]

Web phrase frequency

Our assumptions

- Web phrases are the most reasonable query segments
- More frequent or prominent web phrases are better segments

Web phrase frequency normalization

Our assumptions

- Web phrases are the most reasonable query segments
- More frequent or prominent web phrases are better segments

Our approaches

- Score segmentations based on normalized web frequencies
- But no supervised learning, web retrieval, etc.

Web phrase frequency normalization

Our assumptions

- Web phrases are the most reasonable query segments
- More frequent or prominent web phrases are better segments

Our approaches

- Score segmentations based on normalized web frequencies
- But no supervised learning, web retrieval, etc.

Our web representation

- Collection of the 1- to 5-grams from the 2006 Google index
- Including occurrence frequencies ≥ 40 [Brants and Franz, LDC 2006]

Step 1: Fetch *n*-gram frequencies of potential segments

segment s	$freq(s)$
`new york`	165.4 million
`new york times`	17.5 million
`new york times square`	20 476
`new york times square dance`	0
`york times`	17.6 million
`york times square`	20 561
`york times square dance`	0
`times square`	1.3 million
`times square dance`	104
`square dance`	210 440

Step 2: Frequency normalization $|s|^{|s|} \cdot freq(s)$

segment s	$freq(s)$	$weight(s)$
`new york`	165.4 million	661.6 million
`new york times`	17.5 million	472.5 million
`new york times square`	20 476	5.2 million
`new york times square dance`	0	0
`york times`	17.6 million	70.4 million
`york times square`	20 561	0.5 million
`york times square dance`	0	0
`times square`	1.3 million	5.2 million
`times square dance`	104	2808
`square dance`	210 440	0.8 million

Step 3: Score every segmentation S

Summing up the contained weights

$$score(S) = \sum_{s \in S, |s| \geq 2} weight(s)$$

Remarks: We ignore single keywords.
If a *weight* = 0, then *score* = -1.

Example

score("new york" "times square" dance) = 666.8 million
(661.6 million, 5.2 million)

score("new york times square dance") = -1
(0)

Step 4: Select top segmentation from *score* ranking

rank	segmentation S	$score(S)$
1	"new york" "times square" dance	666.8 million
2	"new york" times "square dance"	662.4 million
⋮	⋮	⋮
5	"new york times" "square dance"	473.3 million
⋮	⋮	⋮
13	new york "times square dance"	2808
14	new york times square dance	0
15	"new york times square dance"	-1
16	new "york times square dance"	-1

Frequencies $|s|^{|s|}$-normalized without any semantics.

Frequencies $|s|^{|s|}$-normalized without any semantics.

More “semantics-aware” normalization?

Wikipedia titles as high quality phrases

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Languages
العربية
বাংলা
Català
Česky

Article | Discussion | Read

Times Square

From Wikipedia, the free encyclopedia
(Redirected from Times square)

For the subway station, see Times Square – 42nd Street (New York City Subway). For other uses, see Times Square (disambiguation).

Times Square is a major commercial intersection in the borough of Manhattan in New York City, at the junction of Broadway and Seventh Avenue and stretching from West 42nd to West 47th Streets. The extended Times Square area, also called the Theatre District, consists of the blocks between Sixth and Eighth Avenues from east to west, and West 40th and West 53rd Streets from south to north, making up the western part of the commercial area of Midtown Manhattan.

Step 1.1: Fetch *n*-gram frequencies of potential segments

segment s	$freq(s)$
`new york`	165.4 million
`new york times`	17.5 million
`new york times square`	20 476
`new york times square dance`	0
`york times`	17.6 million
`york times square`	20 561
`york times square dance`	0
`times square`	1.3 million
`times square dance`	104
`square dance`	210 440

Step 1.2: Check which are Wikipedia titles

segment s	$freq(s)$	Wiki
`new york`	165.4 million	✓
`new york times`	17.5 million	✓
`new york times square`	20 476	-
`new york times square dance`	0	-
`york times`	17.6 million	-
`york times square`	20 561	-
`york times square dance`	0	-
`times square`	1.3 million	✓
`times square dance`	104	-
`square dance`	210 440	✓

Step 2.1: Frequency normalization bonus Wiki titles

segment s	$freq(s)$	Wiki	$weight(s)$
`new york`	165.4 million	✓	165.4 million
`new york times`	17.5 million	✓	165.4 million
`new york times square`	20 476	-	20 476
`new york times square dance`	0	-	0
`york times`	17.6 million	-	17.6 million
`york times square`	20 561	-	20 561
`york times square dance`	0	-	0
`times square`	1.3 million	✓	1.3 million
`times square dance`	104	-	104
`square dance`	210 440	✓	210 440

Step 2.2: Frequency normalization $|s| \cdot weight(s)$

segment s	$freq(s)$	Wiki	$weight(s)$
`new york`	165.4 million	✓	330.8 million
`new york times`	17.5 million	✓	496.2 million
`new york times square`	20 476	-	81 904
`new york times square dance`	0	-	0
`york times`	17.6 million	-	35.2 million
`york times square`	20 561	-	61 683
`york times square dance`	0	-	0
`times square`	1.3 million	✓	2.6 million
`times square dance`	104	-	312
`square dance`	210 440	✓	420 880

Steps 3 & 4: Sum up and select top rank (as before)

rank	trend	segmentation S	$score(S)$
1	⇈	"new york times" "square dance"	496.6 million
2	⇈	"new york times" square dance	496.2 million
3	↓	"new york" "times square" dance	333.4 million
⋮	⋮	⋮	⋮
13	-	new york "times square dance"	312
14	-	new york times square dance	0
15	-	"new york times square dance"	-1
16	-	new "york times square dance"	-1

What about accuracy?

How to measure accuracy of segmentations?

The standard corpus

- 500 queries from the AOL log
- Each segmented by 3 human annotators
- Often used for evaluation [Bergsma and Wang, EMNLP-CoNLL 2007]

The standard accuracy measures

- Query level: ratio of correctly quoted queries
- Segment level: precision/recall of computed segments
- Break level: ratio of correct decisions in-between keywords

How accurate are we? (on the Bergsma-Wang corpus)

Annotator	Accuracy	MI	Bergsma-Wang	Naïve	Wiki-based
Best Match	query	0.583	0.702	0.700	**0.726**
	seg prec	0.693	0.812	0.800	**0.820**
	seg rec	0.697	**0.831**	0.796	0.807
	seg F	0.695	**0.821**	0.798	0.814
	break	0.849	0.899	0.889	**0.900**

Observations

- Wiki-based has best query accuracy
- Bergsma-Wang approach counters with best segment recall

Our new evaluation corpus

Shortcomings of the Bergsma-Wang corpus

- Small and not representative (500 queries, just noun-phrases)
- Some duplicate queries, typos, and encoding errors

Our improved corpus

- 50 000 queries (3–10 keywords) sampled from "filtered" AOL log
- Preserving query frequency and length distribution
- Semi-automatic spell checking (14% of the queries corrected)
- Up to 10 annotators per query via Mechanical Turk

Our new evaluation corpus

Shortcomings of the Bergsma-Wang corpus

- Small and not representative (500 queries, just noun-phrases)
- Some duplicate queries, typos, and encoding errors

Our improved corpus

- 50 000 queries (3–10 keywords) sampled from "filtered" AOL log
- Preserving query frequency and length distribution
- Semi-automatic spell checking (14% of the queries corrected)
- Up to 10 annotators per query via Mechanical Turk

Again: How accurate are we? (on our corpus)

Annotator	Accuracy	MI	Naïve	Wiki-based
Best Match	query	0.598	0.599	**0.616**
	seg prec	0.727	0.736	**0.744**
	seg rec	0.738	0.733	**0.739**
	seg F	0.732	0.734	**0.742**
	break	0.844	0.842	**0.850**

Observations

- Performance drop compared to Bergsma-Wang corpus
- MI is a challenging baseline!

What about efficiency?

How fast do we quote?

System and implementation details

- Standard quad-core PC running Ubuntu 10.04
- Hash tables for *n*-gram frequencies and Wikipedia titles
- Need about 13 GB of RAM [Brants et al., EMNLP-CoNLL 2007]

Throughput

3 000 queries per second

Remark: A load of 1 billion queries per day means 12 000 queries per second.

Almost the end: The take-away messages!

What we have done

Results

- Naïve $|s|^{|s|}$-normalization
- Wikipedia-based normalization
- Simple and fast
- As accurate as state of the art
- Improved test corpus

Future Work

- Ranking-aware accuracy
- Retrieval-aware accuracy

What we have (not) done

Results

- Naïve $|s|^{|s|}$-normalization
- Wikipedia-based normalization
- Simple and fast
- As accurate as state of the art
- Improved test corpus

Future Work

- Ranking-aware accuracy
- Retrieval-aware accuracy

What we have (not) done

Results

- Naïve $|s|^{|s|}$-normalization
- Wikipedia-based normalization
- Simple and fast
- As accurate as state of the art
- Improved test corpus

Future Work

- Ranking-aware accuracy
- Retrieval-aware accuracy

Thank you

☺