Toward Whole-Session Relevance: Exploring Intrinsic Diversity in Web Search

Karthik Raman (Cornell University)
Paul N. Bennett (MSR, Redmond)
Kevyn Collins-Thompson (MSR, Redmond)

Whole-Session Relevance

. . . “snow leopards”

bing snow leopards

129,000 RESULTS Any time

NatGeo page on snow leopards

Snowleopard.org new article

News about snow leopards in Cape May

Snow leopard babies at Boise Zoo

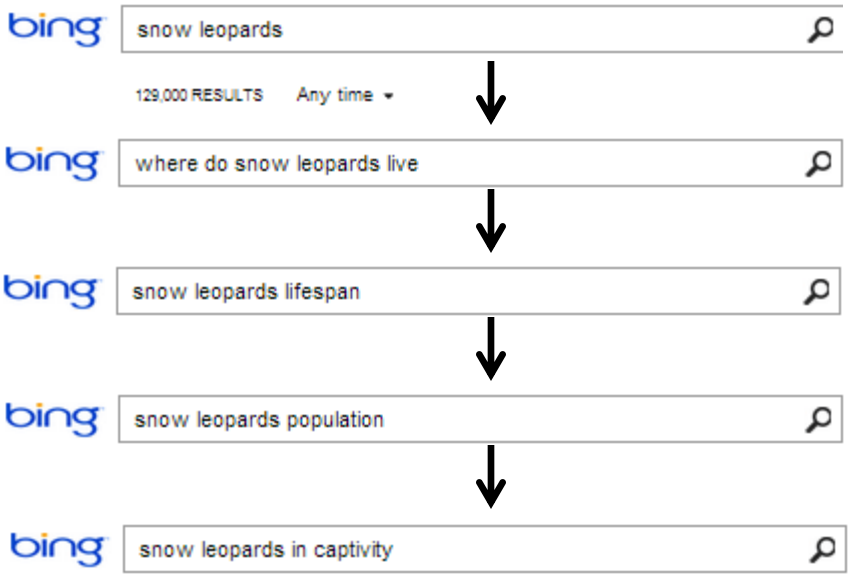
BBC video on snow leopards triplets

- Typical search model :

  Present results maximizing relevance to current query

Whole-Session Relevance

. . . “snow leopards”



- Typical search model :

  Present results maximizing relevance to current query
- Context can improve search.
- **Time** and **user effort** matter! [Smucker&Clarke,2012]
- Instead :

  Present results maximizing relevance to current and future (in-session) queries

Whole-Session Relevance

. . . “snow leopards”

- Typical search model :

  Present results maximizing relevance to current query
- Context can improve search.
- **Time** and **user effort** matter! [Smucker&Clarke,2012]
- Instead :

  Present results maximizing relevance to current and future (in-session) queries
- Satisfy users up-front!
- Pre-fetch apropos content

Intrinsic Diversity

- Traditional (*extrinsic*) diversity:
  - Ambiguity in user intent.
- *Intrinsic* Diversity [Radlinski et al '09]
  - User wants diverse results *i.e.,* diversity intrinsic to need.
  - Single topical intent but diverse across different aspects.
  - Seen in previous example.
- Traditional diversification methods not well-suited:
  - Need to diversify across aspects of a ***single*** intent ***not*** user intents.
  - Observed empirically as well.

Significance of Intrinsic Diversity

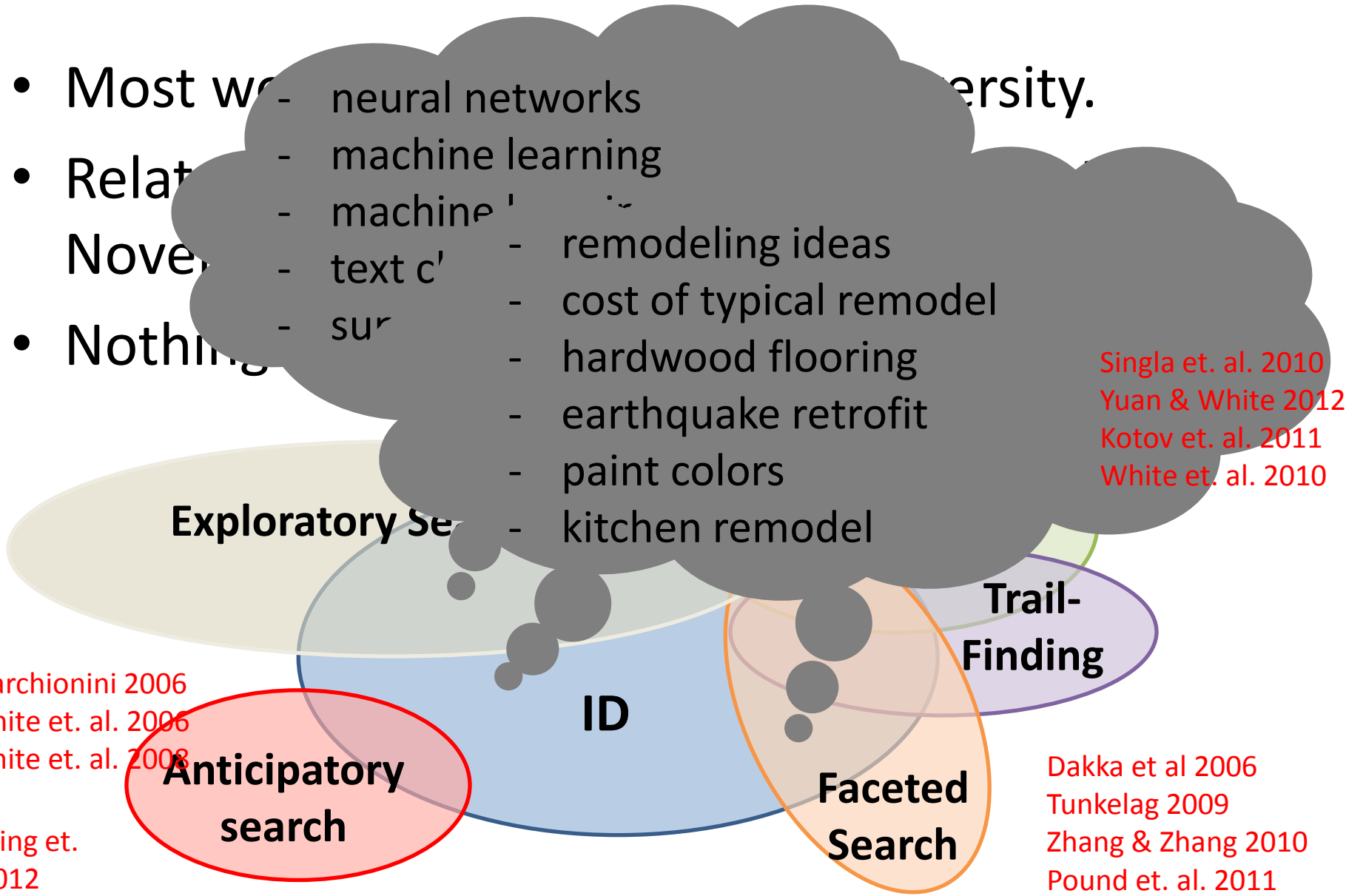
- Bailey et. al. (2012) studi… real world sea… characterize…
  - Require…
- For examp…

- Best Tablet Reade… …ple
- kelly clarkson superbowl performance
- remodeling kitchen
- installing kitchen cabinets
- Installing base cabinets
- how to attach countertop to base cabinets?
- hanging wall cabinets

Que…	…valence …ession)
Information Discov…	14%
Comparing Products …	12%
Finding Facts about a per…	6.9	4.8	3.5%
Learning to perform a task	13	8.5	2.5%

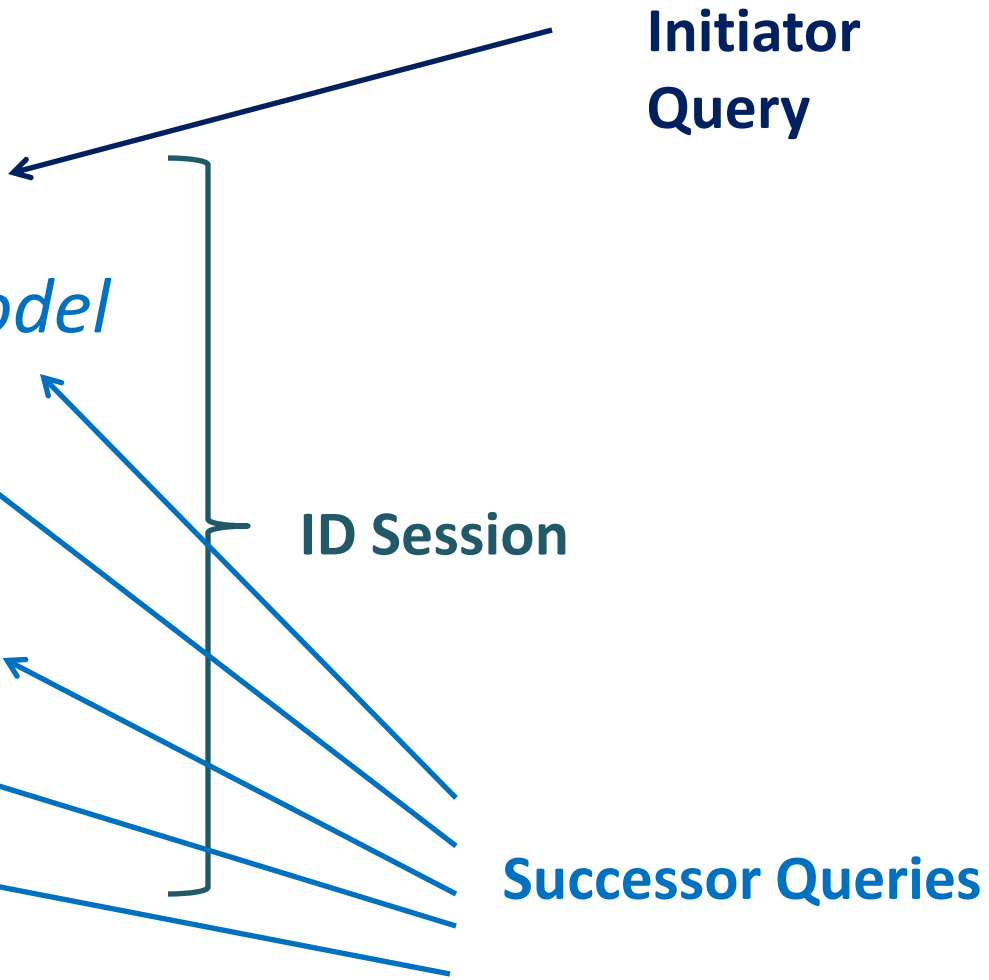
Related Problems

- Most w… …ersity.
- Relat…
  Nove…
- Nothi…



Example ID session

- facebook
- remodeling ideas
- *cost of typical remodel*
- *hardwood flooring*
- cnn news
- *earthquake retrofit*
- *paint colors*
- *kitchen remodel*
- nfl scores
- ----



Our Contributions

1. Mining ID sessions from post-hoc behavioral analysis in search logs.

2. Learning to predict initiator queries of ID sessions.

3. Given initiator query, rank results targeting whole-session relevance and also predict which content to pre-fetch.

Our Contributions

1. Mining ID sessions from post-hoc behavioral analysis in search logs.

2. Learning to predict initiator queries of ID sessions.

3. Given initiator query, rank results targeting whole-session relevance and also predict which content to pre-fetch.

Mining ID sessions from logs

- Would like authentic ID session instances.
- Mine from query logs of a search engine.
- Hypothesize ID Sessions to be:
  1. **Longer:** User explores multiple aspects.
  2. **Topically Coherent:** Aspects should be topically related.
  3. **Diverse in Aspects:** Not just simple reformulations.

- Proposed algorithm is a series of filters.

ID Extraction Algorithm: Key Steps

1. Query Filtering

- facebook
- remodeling ideas
- ideas for remodeling
- cost of typical remodel
- hardwood flooring
- cnn news
- earthquake retrofit
- paint colors
- dublin tourism
- kitchen remodel
- nfl scores

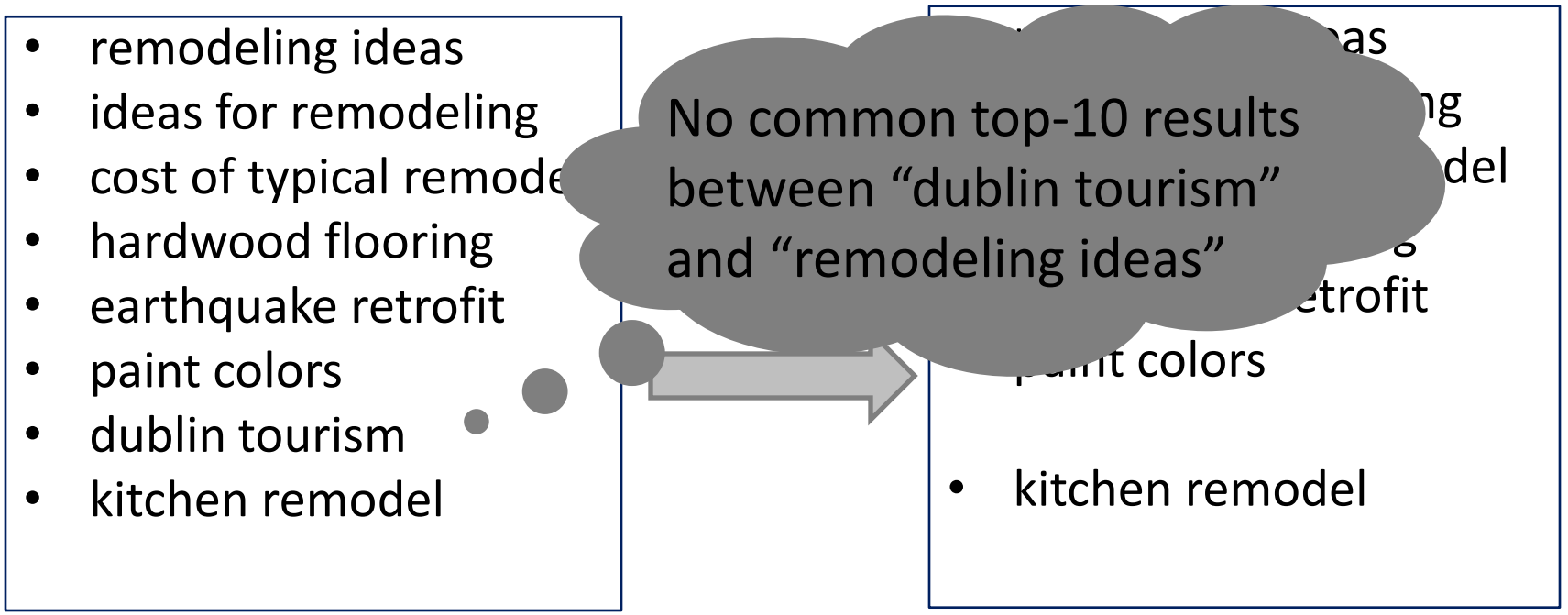
- remodeling ideas
- ideas for remodeling
- cost of typical remodel
- hardwood flooring

- earthquake retrofit
- paint colors
- dublin tourism
- kitchen remodel

- Remove common queries, auto-generated queries, long queries. Collapse duplicates.

ID Extraction Algorithm: Key Steps

2. Ensure topical coherence



- Remove successor queries topically unrelated to initiator.
- >= 1 common result in top 10 (ensures semantic relatedness w/o requiring ontology).

ID Extraction Algorithm: Key Steps

3. Ensure diversity in aspects

- remodeling ideas
- ideas for remodeling
- cost of typical remodel
- hardwood flooring
- earthquake retrofit
- paint colors
- kitchen remod

- remodeling ideas
- cost of typical remodel

Trigram-Cosine Similarity with "remodeling ideas":

- "ideas for remodeling" .693 ✗
- "cost of typical remodel" .292 ✓
- "hardwood flooring" .000 ✓
- "earthquake retrofit" .000 ✓
- "paint colors" .000 ✓
- "kitchen remodel" .371 ✓

….

- Restrict syntacti
  among successo
- Used character-based t milarity.

ID Extraction Algorithm: Key Steps

4. Ensure minimum length

- remodeling ideas
- cost of typical remodel
- hardwood flooring
- earthquake retrofit
- paint colors
- kitchen remodel

>= 2 distinct aspects?

- Ensure minimum number of (syntactically) distinct successor queries *i.e.,* aspect threshold.

Evaluating Extraction

- Previously unstudied problem.
  - Thus quantitatively evaluated by 2 annotators.
- Annotated 150 random sessions:
  - 75 selected by algorithm (as ID) + 75 unselected sessions.

Annotator Agreement	Algorithm Accuracy
79%	73.7% (Prec:73.9%)

- Use this as (noisy) supervision:
  - Sessions selected called ID. Others called regular.
- Given enough data, learner can overcome label noise (if unbiased) [Bartlett et al '04].

Statistics of Extraction Process

- Started with 2 months log data:
  - 51.2 M sessions (comprising 134M queries)
- Running the extraction algorithm leads to 497K sessions (comprising 7M queries)
- Accounts for 1% of sessions but 4.3% of time spent searching.

Our Contributions

1. Mining ID sessions from post-hoc behavioral analysis in search logs.

2. Learning to predict initiator queries of ID sessions.

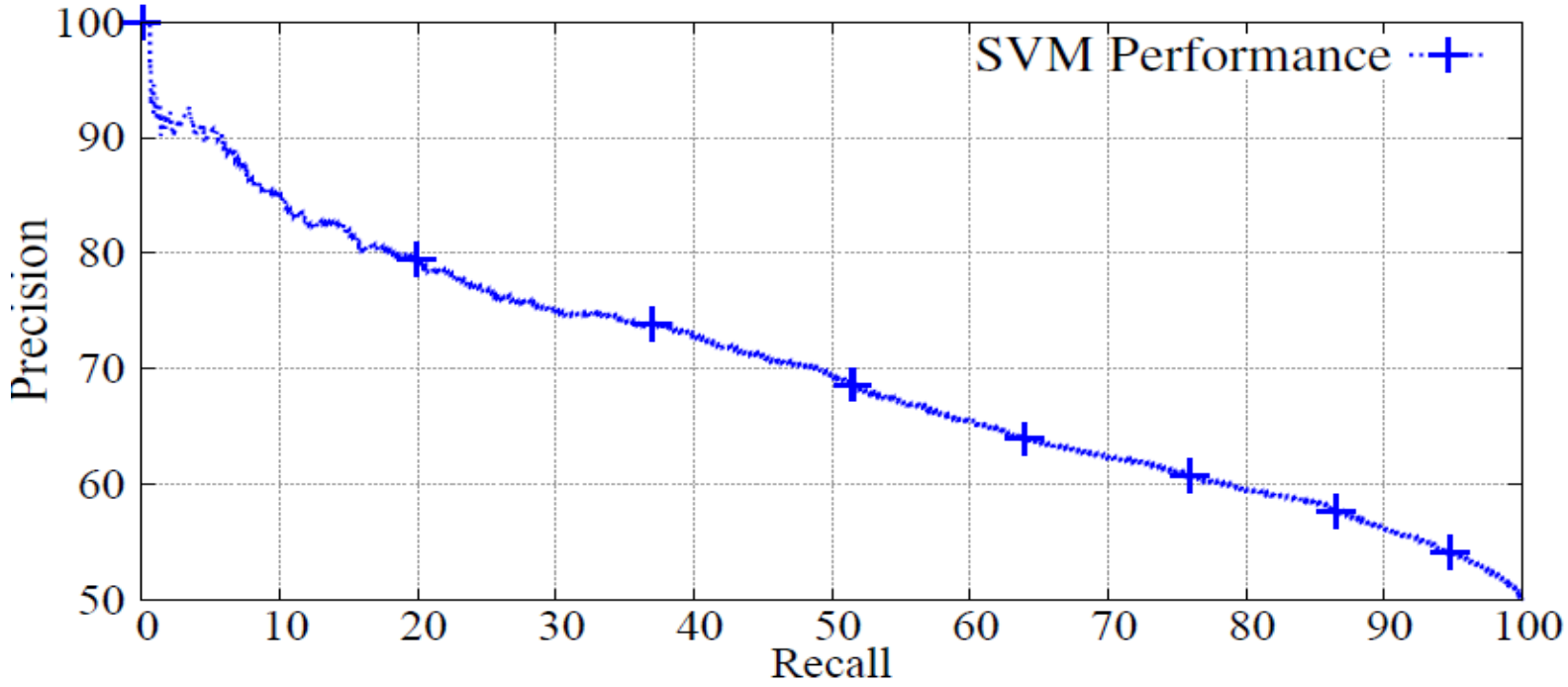
3. Given initiator query, rank results targeting whole-session relevance and also predict which content to pre-fetch.

Predicting ID Initiation

- Can alter retrieval for ID sessions:
  - Example: *Prefetch content/use different ranker ..*
  - Hence **need to identify ID initiation**.
- Given (initiator) query, binary classification problem: Is the session ID or Regular?
- Novel prediction task:
  - New type of query and session being analyzed.

ID Initiation Classification

- Labels produced by extraction algorithm.
- Balanced dataset: 61K unique queries (50K train)
- Used linear SVMs for classification



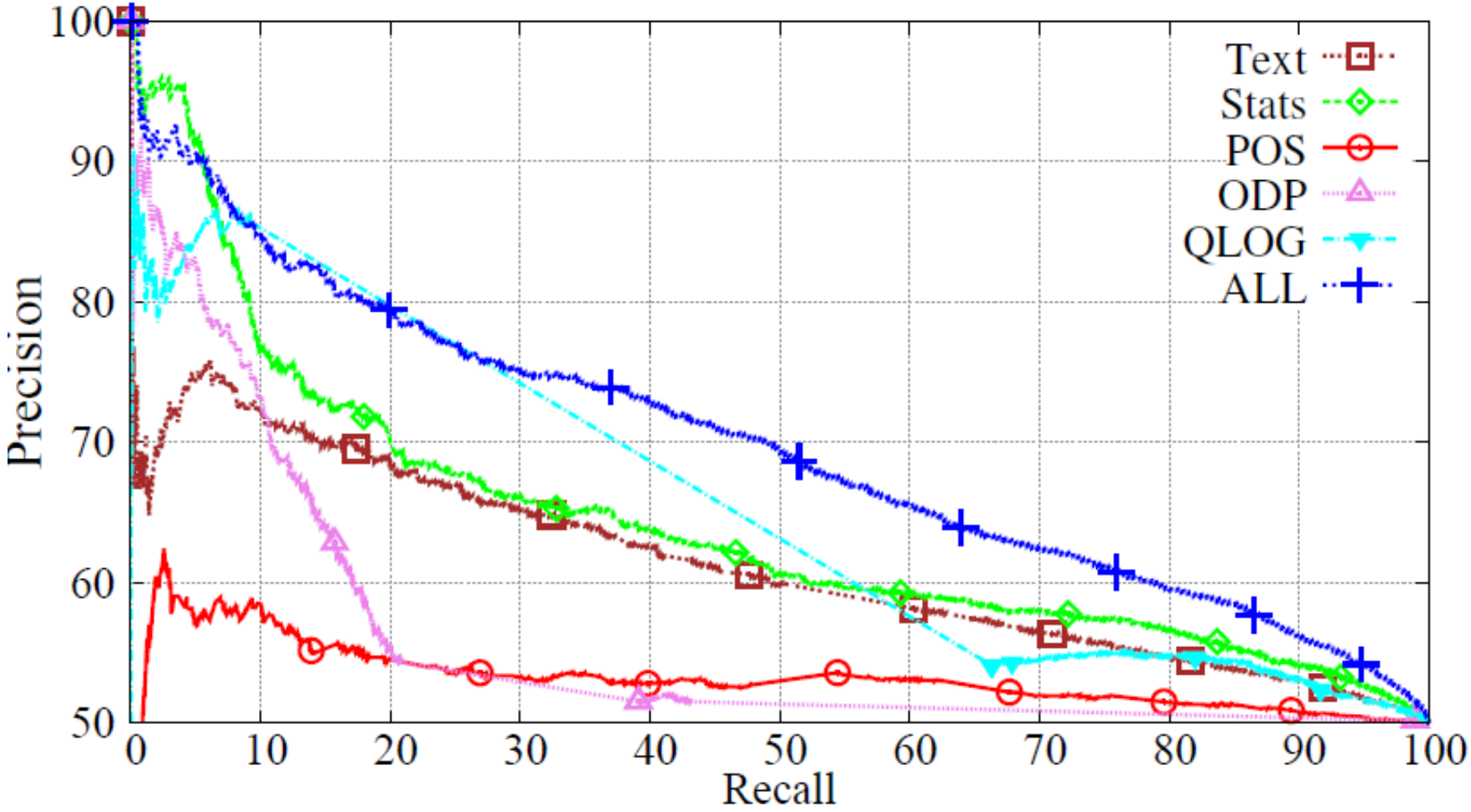
- Can achieve 80% precision@20% recall.

Digging Deeper: ID Initiation Features

- 5 types of features:

TYPE	Description
Textual	B.O.W. (Unigram) counts
Query-Statistics	e.g. # Words
POS	Part-of-speech tag counts
ODP Categories	5 Most probable ODP classes
Query-Log Based Statistics	e.g. Avg. session length

ID Initiation Feature Importance



- Text, Stats and Query-Log features most useful.

Linguistic Characterization of ID Queries

- Measured Log-Odds-Ratio (LOR) of linguistic features:
  - Higher LOR = more pronounced in ID queries.
  - List-like nouns appear more commonly.
  - Broad information-need terms as well.
  - Question words (e.g. *who, what, where*) and proper nouns (e.g. *Kelly Clarkson, Kindle*) quite indicative of being ID.
  - Plural nouns (e.g. *facets, people*) favored to singular nouns (e.g. *table*).

forms	**1.59**
facts	**1.45**
types	**1.25**
ideas	**0.92**
information	**1.64**
manual	**1.18**
Question W	**0.41**
Proper N	**0.4**
Plural N	**0.13**
Singular N	**-0.05**

Our Contributions

1. Mining ID sessions from post-hoc behavioral analysis in search logs.

2. Learning to predict initiator queries of ID sessions.

3. Given initiator query, rank results targeting whole-session relevance and also predict which content to pre-fetch.

Ranking for ID sessions

- Problem: Given initiator query, rerank to maximize whole-session relevance.

- First to jointly satisfy current and future queries.
- Need to identify content to pre-fetch.

- Rank results by associating each with an aspect.
- Candidate pool of aspects generated using related queries.

Ranking Algorithm

- Given query *q*:

Produce ranking d_1, d_2.. (with associated aspects q_1, q_2..)

1. Documents should be relevant to query *q.*
2. Document d_i should be relevant to associated aspect q_i .
3. Aspects should be relevant to ID task initiated by *q*.
4. Aspects should be diverse.

- Objective :

$$\text{argmax}_{(d_1,q_1)\cdots(d_n,q_n)} \sum_{i=1}^{n} \gamma_i \cdot R(d_i|q) \cdot R(d_i|q_i) \cdot e^{\beta Div(q_i,\mathcal{Q})}$$

Breaking Down the Objective - 1

$$\text{argmax}_{(d_1,q_1)\cdots(d_n,q_n)} \sum_{i=1}^{n} \gamma_i \cdot R(d_i|q) \cdot R(d_i|q_i) \cdot e^{\beta Div(q_i,\mathcal{Q})}$$

- Document relevance to query.
- Trained Relevance model (with 21 simple features) using Boosted Trees.

Breaking Down the Objective - 2

$$\text{argmax}_{(d_1,q_1)\cdots(d_n,q_n)} \sum_{i=1}^{n} \gamma_i \cdot R(d_i|q) \cdot R(d_i|q_i) \cdot e^{\beta Div(q_i,\mathcal{Q})}$$

- Document relevance to aspect.
  - *Represents/Summarizes* the aspect.

- Can be estimated with same relevance model *R*

Breaking Down the Objective - 3

$$\text{argmax}_{(d_1,q_1)\cdots(d_n,q_n)} \sum_{i=1}^{n} \gamma_i \cdot R(d_i|q) \cdot R(d_i|q_i) \cdot e^{\beta Div(q_i,\mathcal{Q})}$$

- Aspect Diversity + Topical Relevance.
- MMR-like objective

$$Div(q_i, \mathcal{Q}) = \lambda \cdot Sim(q_i, Snip(q)) - (1-\lambda) \max_{j<i} Sim(Snip(q_i), Snip(q_j)).$$

- Submodular Objective:
  - Optimize using efficient greedy algorithm.
  - Constant-factor approximation.

Performance on Search Log Data

- Measured performance as ratio (to baseline ranker)
- Baseline is the commercial search engine service.
- Relevance-based: ranking with $R(d|q)$.
- ID Session SAT clicks used as relevant docs.

Method	PREC			MAP			NDCG		
	@1	@3	@10	@1	@3	@10	@1	@3	@10
Relevance-Based	1.00	0.94	0.97	1.00	0.97	0.98	1.00	0.97	0.99
Proposed Method	**1.10**	**1.09**	**1.09**	**1.10**	**1.10**	**1.10**	**1.09**	**1.10**	**1.11**

Other Findings on Search Log Data

- Robust: Very few sessions drastically hurt.
- Similar performance on using sessions *classified* as ID (by the SVM)
- Even more improvements (30-40%) on using interactivity (based on simple user model).
- A good set of aspects can greatly help: 40-50% increase w/o interactivity; 80-120% with it.

Performance on TREC data

- Also ran experiments using public dataset:
  - TREC 2011 Session data
  - 63/76 annotated as ID.
  - Absolute (not relative) performance values reported.

METHOD	Pr@1	DCG@1	DCG@3
Baseline	0.58	0.84	2.13
Proposed	0.71	1.39	2.41

Contributions Recap

- First study of Intrinsic Diversity for Web Search.
- Method to mine ID examples from logs.
- Characterized and predicted ID initiation.
- Presented ranking algorithm for ID sessions maximizing whole-session relevance.

Toward Whole-Session Relevance

- Retrieval quality can be directly improved to reduce time spent manually querying aspects.
- Presented results can serve as an easy way of summarizing aspects.
- Structuring results to enable users to interactively explore aspects is a step towards this goal.

THANK YOU!

QUESTIONS?

Thanks to SIGIR for their generous SIGIR Travel Grant.

THANK YOU!

QUESTIONS?

Thanks to SIGIR for their generous SIGIR Travel Grant.

BACKUP SLIDES

Scope and Applicability

- Clearly not feasible for all kinds of sessions!
- So what can we handle?
  - Breadth-oriented sessions.
  - Exploratory sessions.
  - Comparative sessions.
- **Intrinsic Diversity**:
  - Underlying information need tends to be of one of the above forms.

ID Initiation Classification

- Balanced dataset: 61K unique queries (50K train)
- Used linear SVMs for classification
- 5 types of features:

TYPE	Description	# of Feat.	Coverage
Text	B.O.W. (Unigram) counts	44k	100%
Stats	e.g. # Words	10	81%
POS	Part-of-speech tag counts	37	100%
ODP	5 Most probable ODP classes	219	25%
QLOG	e.g. Avg. session length	55	44%

Examples: Misclassified as ID

- Precision Level indicates where on the spectrum it lies.

Precision	Queries
>90	adobe flash player 10 activex bing maps live satellite aerial maps how old is my house
~90	port orchard jail roster was is form 5498 java file reader example free ringtones downloads
~80	life lift top pit masters in the state promag 53 user manual pky properties llc nj

Examples: Misclassified as Regular

- Precision Level indicates where on the spectrum it lies.

Precision	Queries
~65	assisted living coppell texas and sandy lake www jsu com "visions dpsnc net res mychecks"
~62	uab graduation announcements examples of reception invitations ndc mla handbook online
~60	wedding notary public seffner fl redshedtoys sunrealty uranium natural state

Feature-Wise Errors

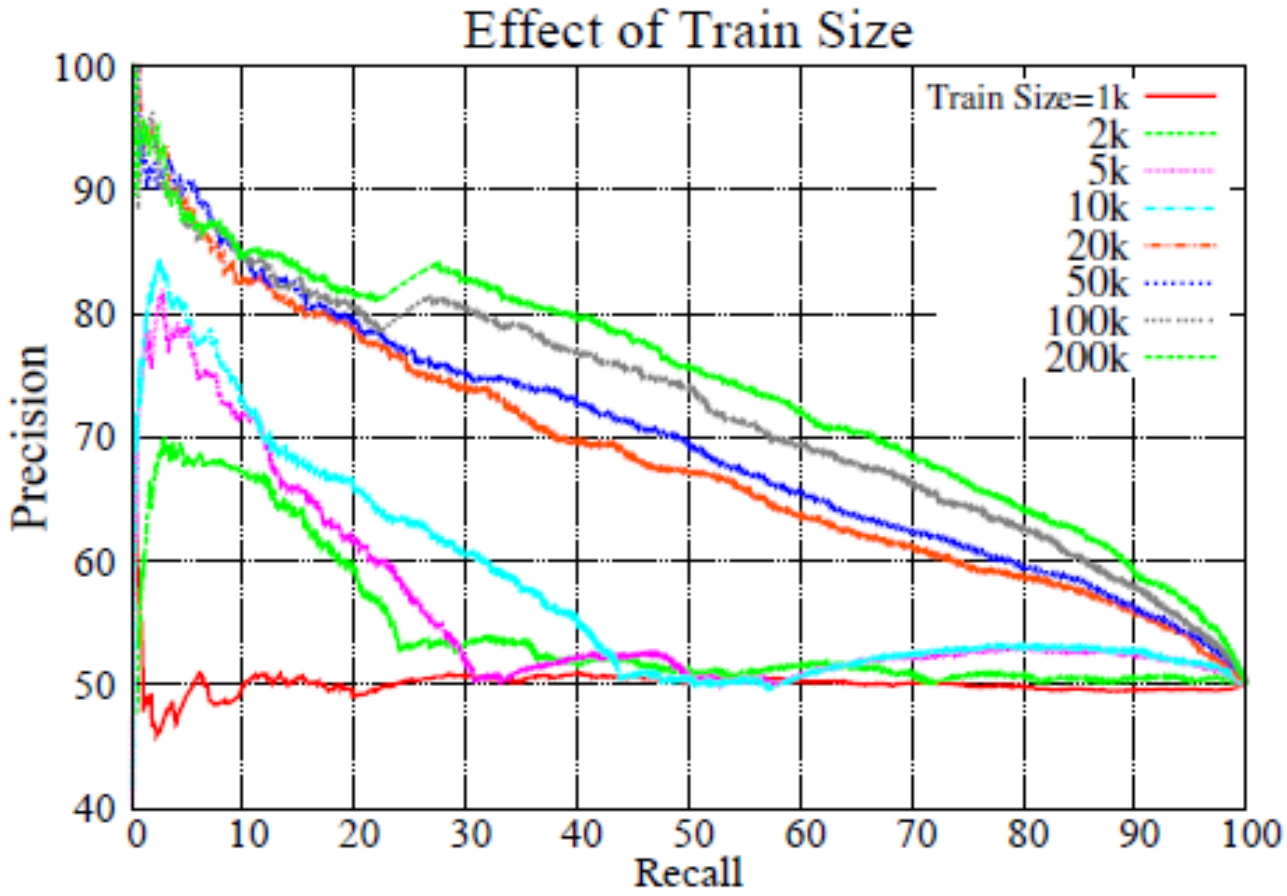
- Misclassifications for different feature sets.

Table VII. *Regular* queries at high precision levels for different feature sets.

Feature	Examples
Text	[nebraska state income tax return forms] [what are different types of plaids] [free diet plans]
Stats	[how old is my house] [live satellite aerial maps] [this is how i dew it] [bing maps]
ODP	[texas press association]
QLOG	[entertainment in kansas] [im a g lyrics] [winchester gun safes] [juegos de futbol]

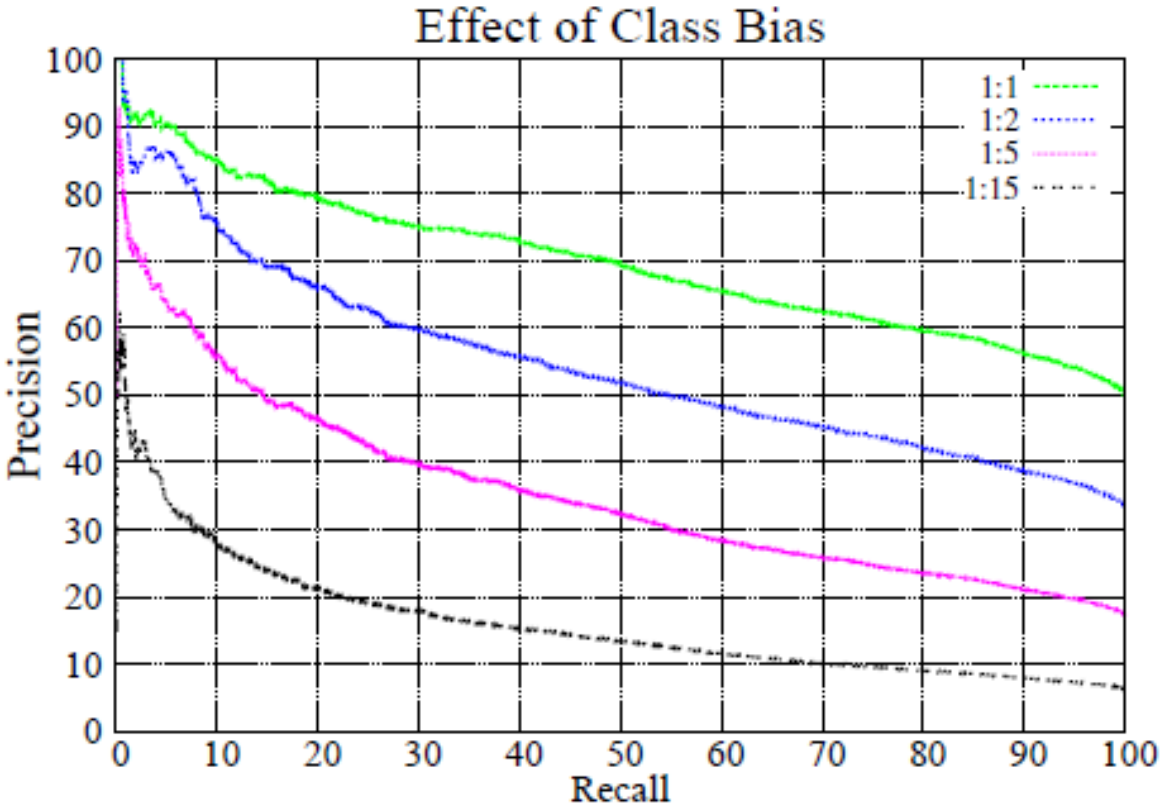
Effect of Training Size

- More the data, the better.



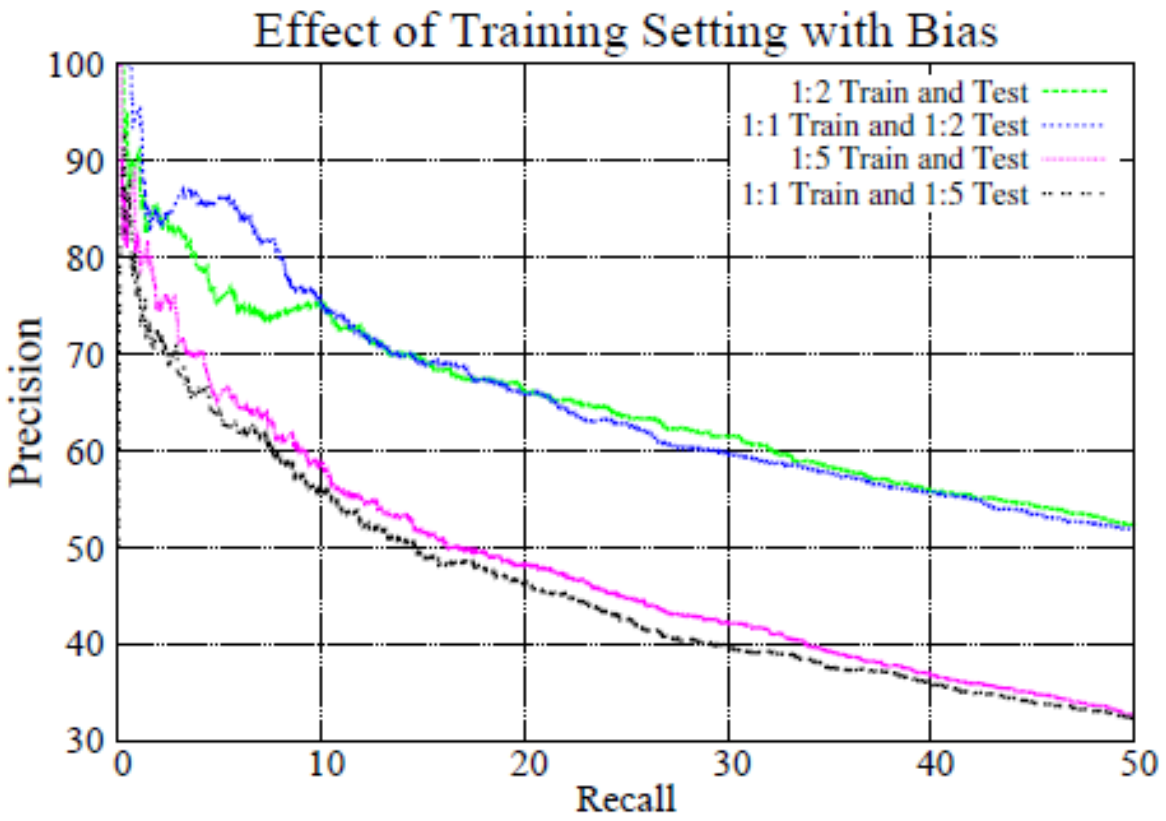
Effect of class bias

- No longer balanced dataset.



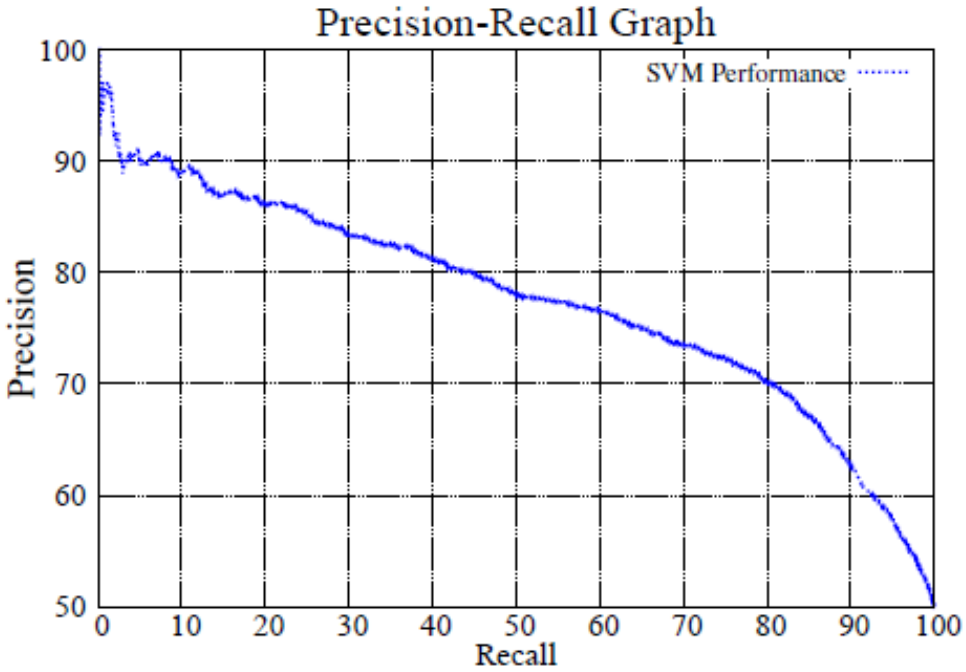
Training effect of class bias

- No longer balanced dataset.
- Train and Test have different class ratios.



All-Query Classification

- Learning to classify if ANY query in a session is part of ID session or not.
- Can be used for identifying when ID is over (or off-topic query).



bing MS Beta

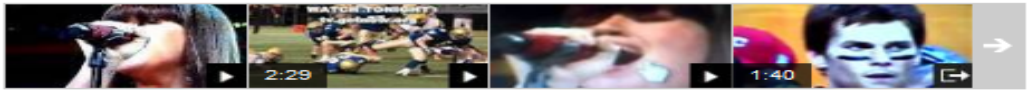
kelly clarkson super bowl performance

30,300,000 RESULTS Any time

Kelly Clarkson keeps national anthem **Super Bowl performance** ...
Entertainment Weekly Online
Feb 05, 2012 · There were no flubbed lyrics, and no screeching, when **Kelly Clarkson** sang the national anthem at **Super Bowl** XLVI on Sunday. In fact, **Clarkson** kept ...

Kelly Clarkson's **Super Bowl Performance** Truly Idol-Worship ...
YAHOO!
photo: Ezra Shaw/ Getty ImagesLast year at the **Super Bowl**, Christina Aguilera's national anthem **performance** was a disaster on an international scale. And this year ...

Videos of **kelly clarkson super bowl performance**
bing.com/videos



Kelly Clarkson National Anthem YouTube
Kelly Clarkson Super Bowl 201... Dailymotion
Kelly Clarkson's Super Bowl YouTube
Kelly Clarkson Canta el Himno metatube

Kelly Clarkson's **Super Bowl Performance** Truly Idol-Worship ...
omg! on Yahoo!
Last year at the **Super Bowl**, Christina Aguilera's national anthem **performance** was a disaster on an international scale. But this year, Xtina's new "The Voice ...

Super Bowl | The Music Mix | EW.com
music-mix.ew.com/category/**super-bowl**
Kelly Clarkson keeps it quick and simple for her **Super Bowl performance** of the national anthem ... no screeching, when **Kelly Clarkson** sang the national anthem at **Super Bowl** ...

Images of **kelly clarkson super bowl performance**
bing.com/images

News about **kelly clarkson super bowl performance**
bing.com/news

Kelly Clarkson Belts Out 'Dark Side' + 'Stronger' at 2012 MuchMusic Video Awards
Kelly Clarkson may have lighter hair now, but she still has a 'Dark Side' — and she brought it, as well as 'Stronger,' to the 2012 MuchMusic Video Awards... Popcrush · 9 days ago
The **Super Bowl** Is Now Everyone's Game Huffington Post
Duets' Goes to the Movies: TV Recap Wall Street Journal

Super Bowl 2012: **Kelly Clarkson** Nails National Anthem (Video ...
www.hollywoodreporter.com/earshot/**super-bowl**-2012-**kelly-clarkson**...
Super Bowl 2012: **Kelly Clarkson** Nails National Anthem (Video) The Grammy-winning diva performs a ... Prior to her **Super Bowl performance**, country music power ...

RELATED SEARCHES
- Kelly Clarkson Super Bowl **2012**
- Kelly Clarkson Super Bowl **Time**
- Kelly Clarkson Super Bowl **Length**
- Kelly Clarkson Super Bowl **Outfit**
- Kelly Clarkson **Singing** Super Bowl
- Kelly Clarkson **National Anthem Video**
- Clarkson **National Anthem**
- Super Bowl **Kelley** Clarkson

bing MS Beta

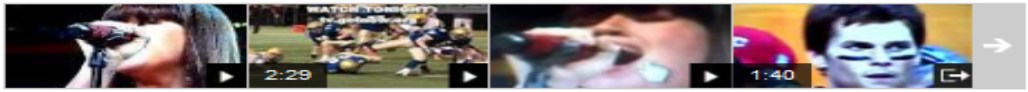
kelly clarkson super bowl performance

30,300,000 RESULTS Any time

Kelly Clarkson keeps national anthem **Super Bowl performance** ...
Entertainment Weekly Online
Feb 05, 2012 · There were no flubbed lyrics, and no screeching, when **Kelly Clarkson** sang the national anthem at **Super Bowl** XLVI on Sunday. In fact, **Clarkson** kept ...

Kelly Clarkson's **Super Bowl Performance** Truly Idol-Worship ...
YAHOO!
photo: Ezra Shaw/ Getty ImagesLast year at the **Super Bowl**, Christina Aguilera's national anthem **performance** was a disaster on an international scale. And this year ...

Videos of **kelly clarkson super bowl performance**
bing.com/videos



Kelly Clarkson National Anthem YouTube | Kelly Clarkson Super Bowl 201... Dailymotion | Kelly Clarkson's Super Bowl YouTube | Kelly Clarkson Canta el Himno metatube

Kelly Clarkson Facts - I AM FAN! Alicia Keys · Jay-Z · Lil ...
www.iamfan.com/~**kelly_clarkson**/**kelly-clarkson-facts**.htm
Kelly Clarkson trivia - interesting **facts about Kelly Clarkson**

Facts About Kelly Clarkson

List of **awards** and nominations received by **Kelly Clarkson** ...
en.wikipedia.org/wiki/List_of_**Kelly_Clarkson_awards**
This is a list of **awards** that American singer-songwriter **Kelly Clarkson** has received throughout her career, which started following her coronation as the first ...

kelly clarkson awards

Images of **kelly clarkson super bowl performance**
bing.com/images

News about **kelly clarkson super bowl performance**
bing.com/news

Kelly Clarkson Belts Out 'Dark Side' + 'Stronger' at 2012 MuchMusic Video Awards
Kelly Clarkson may have lighter hair now, but she still has a 'Dark Side' — and she brought it, as well as 'Stronger,' to the 2012 MuchMusic Video Awards... Popcrush · 9 days ago
The **Super Bowl** Is Now Everyone's Game Huffington Post
Duets' Goes to the Movies: TV Recap Wall Street Journal

Kelly Clarkson makes her MMVA return with a flawless **performance** ...
blog.muchmusic.com/**kelly-clarkson**...return-with-a-flawless-**performance**
During last night's MMVAs, **Kelly Clarkson** blew the crowd away with her double **performance** of Dark Side and Stronger. **Performing** her motivational tracks from

kelly clarkson performance

RELATED SEARCHES

Training Relevance Function

- Used 20k queries.
- Optimized for NDCG@5.

Query	Length
Website	Log(PageRank)
Baseline Ranker	Reciprocal Rank (if in top 10)
URL	Length # of Query Terms Covered Fraction of Query Covered TF Cosine sim LM Score(KLD) Jaccard Boolean AND Match Boolean OR Match
Anchor (Weighted)	Same as URL
Anchor (Unweighted)	TF-Cosine Sim KLD Score