Addressing the Challenges of Underspecification in Web Search

Michael Welch
mjwelch@cs.ucla.edu

Why study Web search?

- **Search engines have enormous reach**
  - **Nearly 1 billion queries globally each day**
- **Search engines drive online advertising market**
  - **Google: $6.5 billion advertising revenue for Q2-2010**
- **User satisfaction is essential for market share**
  - **Profit depends on traffic**

Challenges of Underspecification

- **Underspecification causes several problems for search engines**
- **Underspecified user queries**
  - **What can the search engine do about implicit or ambiguous user intent?**
- **Underspecified content**
  - **How can the search engine determine the keywords from sparse, incomplete, unstructured data?**

Contextualization

- **Find more relevant results based on metadata**
  - **How do we know when metadata is important?**
- **We study identifying geo-*localizable* queries**
  - **Queries where user's location (e.g. city) is relevant**
- **Can significantly improve relevance to the user**
  - **Higher clickthrough rates, happier users**
  - **Relevant context for the keywords, higher ad prices**

Search Diversification

- **Queries are often ambiguous**
  - **Difficult for the search engine to know which aspect the user has in mind**
- **Top results often only cover a few aspects**
  - **Users interested in other meanings are unsatisfied**
- **How can a search engine improve their experience?**
  - **Cover a broader range of interpretations**
  - **Without diminishing quality for most currently "happy" users**

Underspecified Content

- **Content can be short, sparse, or incomplete**
  - **Particularly in the case of videos**
- **Difficult to determine the keywords**
  - **Search and ad matching rely on relevant keywords**
- **How can the search engine find meaningful keywords from the content?**
  - **Which methods work best, and under what conditions?**

Outline

- Identifying localizable queries
- Search result diversity
- Generating keywords for video

Outline

- **Identifying localizable queries**
- Search result diversity
- Generating keywords for video

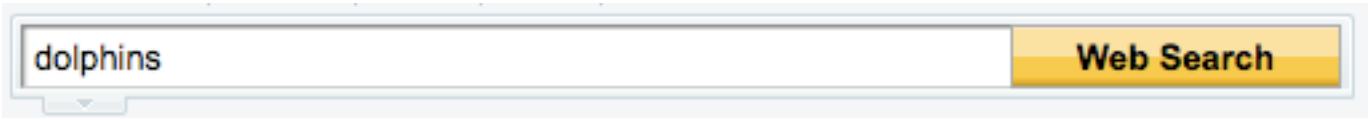
Identifying Localizable Queries

- **Approximately 16% of queries are implicitly geo-localizable [WC08]**
  - **Proposed a framework for automatically identifying these queries**
- **Generated candidate queries from query log**
- **Established distinguishing features**
- **Evaluated well known supervised classifiers on precision and recall**
- **Achieved 94% precision using voting classifier**

Outline

- Identifying localizable queries
- **Search result diversity**
- Generating keywords for video

Search Result Diversity for Informational Queries



Also try: miami **dolphins**, pictures of **dolphins**, More...

Dolphins - Image Results



More **dolphins** images

Yahoo! Shortcut - About

Dolphin - Wikipedia, the free encyclopedia

Origin of the name | Taxonomy | Evolution and anatomy | Behaviour

Dolphins are marine mammals that are closely related to whales and porpoises. There are almost forty species of **dolphin** in seventeen genera. They vary in size from 1.2 m (4 ft) and 40 kg (90 lb), up to 9.5 m (30 ft) and 10 tonnes . They are found worldwide, mostly in...

en.wikipedia.org/wiki/Dolphin - 125k - Cached

Swim with the **dolphins** at Dolphin Research Center Marathon FL, Dolphin ...

A Florida nonprofit education and research facility, home to a family of Atlantic Bottlenose **Dolphins** and California Sea Lions. Offers educational programs that ...

www.dolphins.org - Cached

DOLPHINS

All **dolphins** are toothed whales belonging to the sub-order, odontocetes, of the ... In addition, although the terms **dolphins** and porpoises are often used ...

www.earthtrust.org/wlcurric/**dolphins**.html - Cached

Miami **Dolphins**

Official site of the Miami **Dolphins**. Includes schedule, news, multimedia, photos, player information, statistics, team store, tickets, and more.

www.miamidolphins.com - 1289k

MiamiDolphins.com - Official Website of the Miami **Dolphins**

www.miamidolphins.com/newsite/index.asp - Cached

Dolphins and Man.Equals?

Just how intelligent are **dolphins**? Can humans understand dolphin intelligence? ... Apparently there is something quite impressive about **Dolphins**. ...

www.littletownmart.com/dolphins - Cached

Bottlenose Dolphin - Wikipedia

Description | Taxonomy | Behavior | Intelligence

Bottlenose **dolphins**, the genus Tursiops, are the most common and well-known members of the family Delphinidae, the family of oceanic **dolphins**. Recent molecular studies show the genus contains two species, the Common...

en.wikipedia.org/wiki/Bottlenose_Dolphin - 266k - Cached

(Lack of) Diversity in Results

- **In the top 10 results from a search engine:**
  - 8 are about the mammal
  - 1 is for the NFL team (rank 5)
  - 1 is for an IMAX movie about the mammals (rank 8)
- **What about the other interpretations?**
  - Users interested in them will be dissatisfied

Motivational Questions

- **Are ambiguous queries really a problem?**
  - **16% of Web queries are ambiguous [SLN09]**
- **How many relevant results do users want?**
  - **Did we need to show 8 pages about the mammal?**
  - **Is one page enough? Two pages? Three?**
- **Can we better allocate the top *n* results to cover a more diverse set of subtopics?**
  - **While maintaining user satisfaction for the common subtopics**

Taxonomic Refinement (Related Work)

- **Categorize documents into topic hierarchy**
  - User disambiguates their intent by selecting the subtopic explicitly
  - Open Directory Project
  - Yippy.com (Clusty), Vivisimo, Carrot^2
- **How do you automatically (and accurately) cluster the Web?**
  - There will be incorrectly classified documents
  - Users expect to be rewarded for their extra work

Search Personalization (Related Work)

- **Given a user profile or browsing history, determine the most probable subtopic**
  - **Return documents for that subtopic**
  - **Modeling user profiles in a taxonomy [PG99, LYM02]**
- **May fail due to**
  - **Missing or incomplete user profiles**
  - **Users having diverse or changing interests**
- **Privacy concerns**

Content Based Diversity (Related Work)

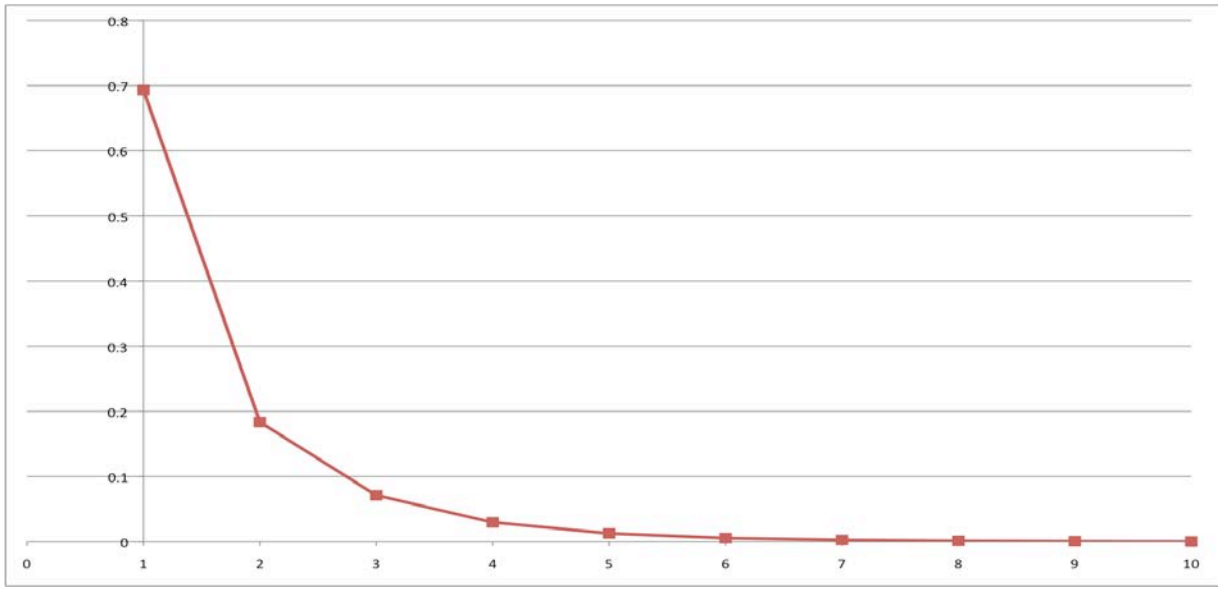
- **Content and language modeling based approaches**
  - **Maximal marginal relevance [CG98]**
  - **Encourage novelty, penalize redundancy [ZCL03]**
  - **Bayesian language modeling [CK06]**
  - **Portfolio theory and managing risk [ZWT09, WZ09]**
- **Diversity as a side effect of novelty**
- **No explicit knowledge of document categorization or user intent**
  - **No way to prioritize the subtopics**

Hybrid Approaches (Related Work)

- **Assume known set of subtopics**
  - **Probabilistic document classifications**
  - **Probabilistic measures of user intent**
- **Return linear list of results aggregated from multiple subtopics**
- **Most existing work assumes a *single* relevant document is sufficient**
  - **Users often require more than one relevant result (e.g. for informational queries)**

Is One Relevant Document Enough?

- One page from the "correct" subtopic may not satisfy every user
- Informational queries typically result in multiple clicks [LLC05]



Our Model for Ambiguous Queries

- **User queries for topic *T* with subtopics $T_1 \ldots T_m$**
- **User has some number of pages *J* that they want to see for their subtopic**
  - **Click on *J* relevant pages if they are available**
    - **Clicks on fewer if less than *J* pages are relevant**
- **Probability of how many pages a user needs**
  - **User *U* wants *J* relevant pages with *Pr(J|U)***

Our Model (cont.)

- **Probabilistic user intent in subtopics**
  - Most users interested in a single subtopic
  - User *U* interested in subtopic T_i with $Pr(T_i|U)$
- **Probabilistic document categorization**
  - Most documents belong to a single subtopic
  - Document *D* belongs to subtopic T_i with $Pr(T_i|D)$

Our Approach for Diversification

- **Model the expected user satisfaction with a returned set of documents**
  - **Optimize document selection for that model**
- **How do we measure user satisfaction?**
  - **Binary "happy or not" isn't an adequate model**
  - **Measure the expected number of *hits***
    - **Hit: a click on a relevant document**
- **We'll start with two simplifications**
  - **Perfect knowledge of user intent**
  - **Perfect document classification**

Perfect Knowledge of User Intent

- **Assume we know which subtopic T_i the user is interested in**

$$E(R) = \sum_{j=1}^{n} \Pr(J = j|U) \sum_{k=1}^{n} \Pr(K_i = k|R) \min(j, k)$$

- **K_i is the *probabilistic* number of documents shown from subtopic T_i**
- **Solution is fairly straightforward**
  - **Choose the documents with highest probability of satisfying T_i**

Perfect Document Classification

- **Now, instead assume we know the correct subtopic for each document**

$$E(R) = \sum_{j=1}^{n} \sum_{i=1}^{m} \Pr(T_i|U) \Pr(J = j|U) \min(j, K_i)$$

- **User is shown K_i pages from subtopic T_i**
- **How many pages should we show from each subtopic T_i?**

Choosing Optimal K_i Values

- **Selecting *n* documents from *m* topics:** $\binom{n+m-1}{n}$
- **Lemma (proof given in dissertation)**
  - **Label subtopics $T_1 \ldots T_m$ such that $Pr(T_1|U) \geq Pr(T_2|U) \geq \ldots Pr(T_m|U)$**
  - **Optimal solution has property $K_1 \geq K_2 \geq \ldots K_m$**
- **Reduces combinations significantly**
  - **Relatively simple to enumerate and test the possible combinations, but we can avoid this in practice**
  - **Combine with *Pr(J|U)* for greedy approach**

KnownClassification Algorithm

- **Start with $K_1 = K_2 = \ldots = K_m = 0$**
- **Choose next subtopic *i* which gives the maximum additional benefit**
  - $i \leftarrow ARGMAX[\ Pr(T_i|U) \times Pr(K_i+1|U)\ ]$
- **Increment K_i**
  - $K_i \leftarrow K_i + 1$
- **Choose next document from subtopic T_i**
  - **e.g. using original search engine ranking function(s)**

Complete Model

- **Given all three probability distributions, we define the expected hits as:**

$$E(R) = \sum_{j=1}^{n} \sum_{i=1}^{m} \Pr(T_i|U) \Pr(J = j|U) \sum_{k=1}^{n} \Pr(K_i = k|R) \min(j, k)$$

- **How to maximize this equation efficiently?**
  - **Take a greedy approach**

Diversity-IQ Algorithm

- ***Start with empty result set R = Ø***
- **Successively choose documents from *D* which give the maximum increase in expected hits**
  - ***d ← ARGMAX[ΔE(d|R,D)]***
- **ΔE computation in O(|R| • |D| • |m|)**
  - **Implement using a greedy approach**
- **Total complexity is polynomial**
  - **$O(n^2 \bullet |D| \bullet |m|)$**

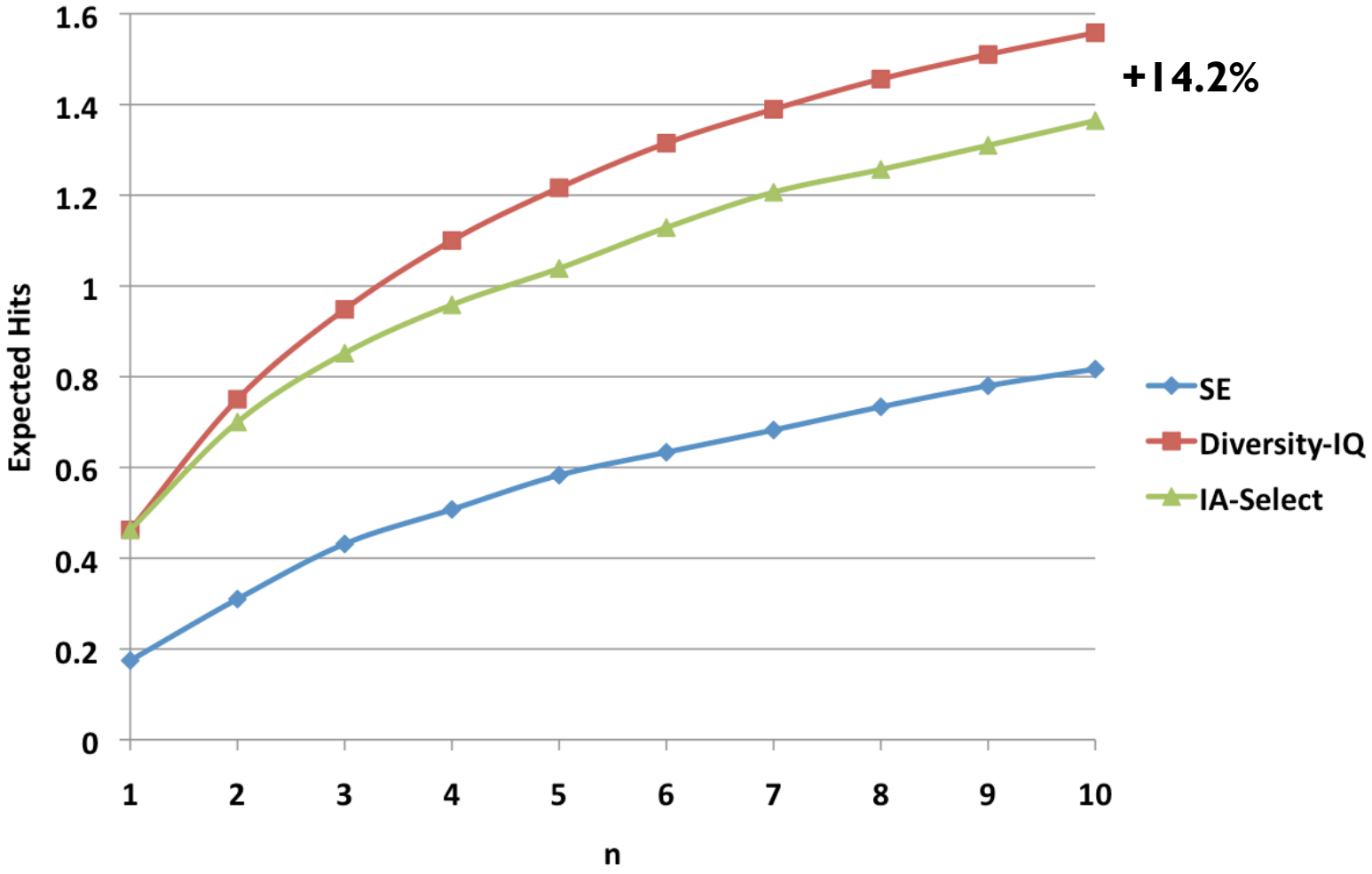
Evaluating *Diversity-IQ*

- **Generated set of 50 ambiguous test queries from Web query log**
- **Extracted subtopic categories from Wikipedia**
  - **Issued each subtopic title as query to search engine and merged top 200 results to form document set**
- **Compared with two other ranking strategies**
  - **Original search engine ranking**
  - **Ranking generated by *IA-Select* [AGH09]**
- **Focused on performance of the top 10 results**

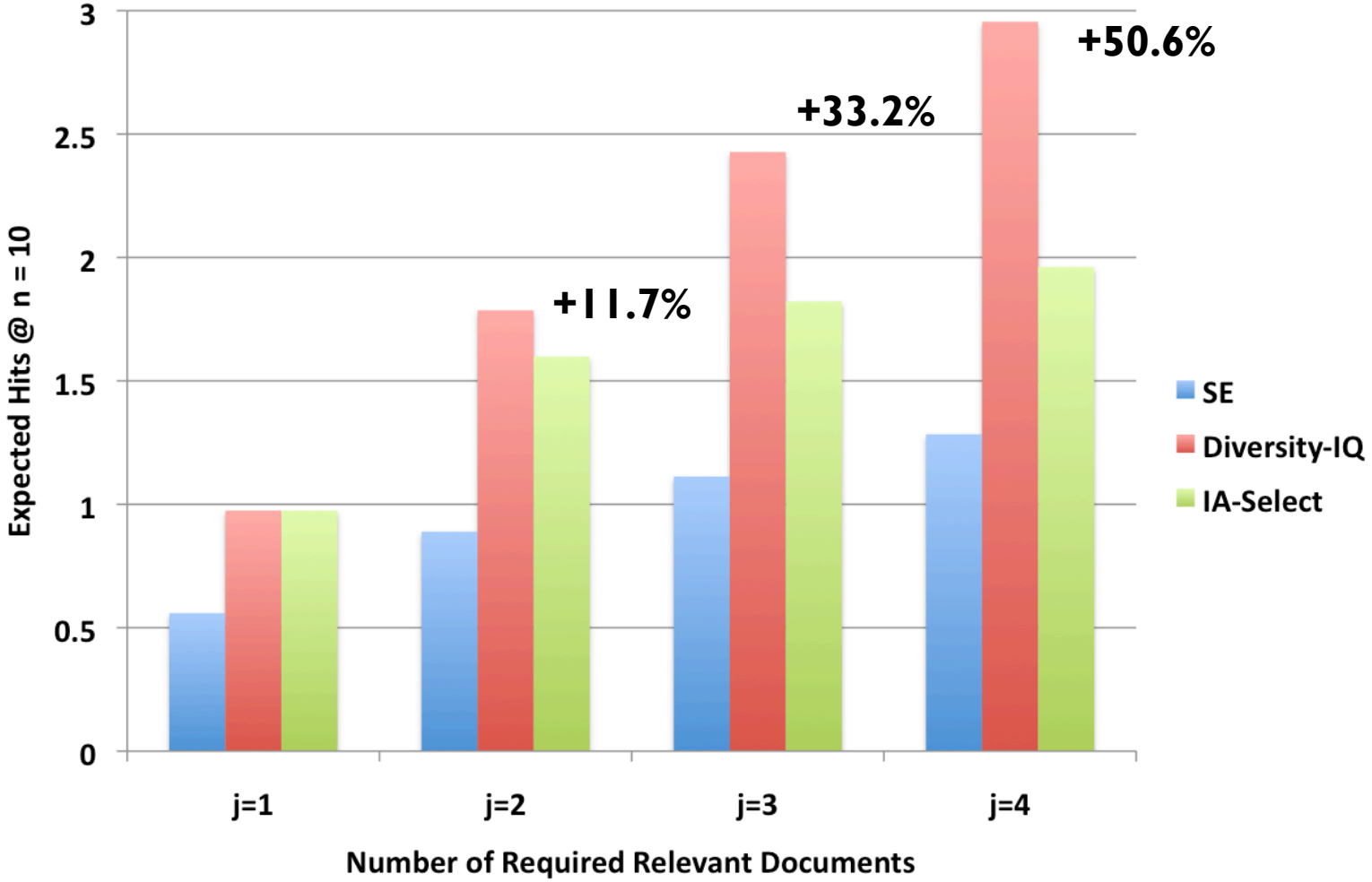
Probability Distributions for Evaluations

- Algorithm needs 3 probability distributions
- Page requirements $Pr(J|U)$
  - Geometric series $Pr(J=j|U) = 2^{-j}$
    - Click log underestimates (e.g. contains navigational)
- User intent $Pr(T_i|U)$
  - Mechanical Turk survey
- Document classification
  - Latent Dirichlet Allocation
    - Used resulting Θ document-topic distribution

Expected Hits



Expected Hits (varying *Pr(J|U)*)



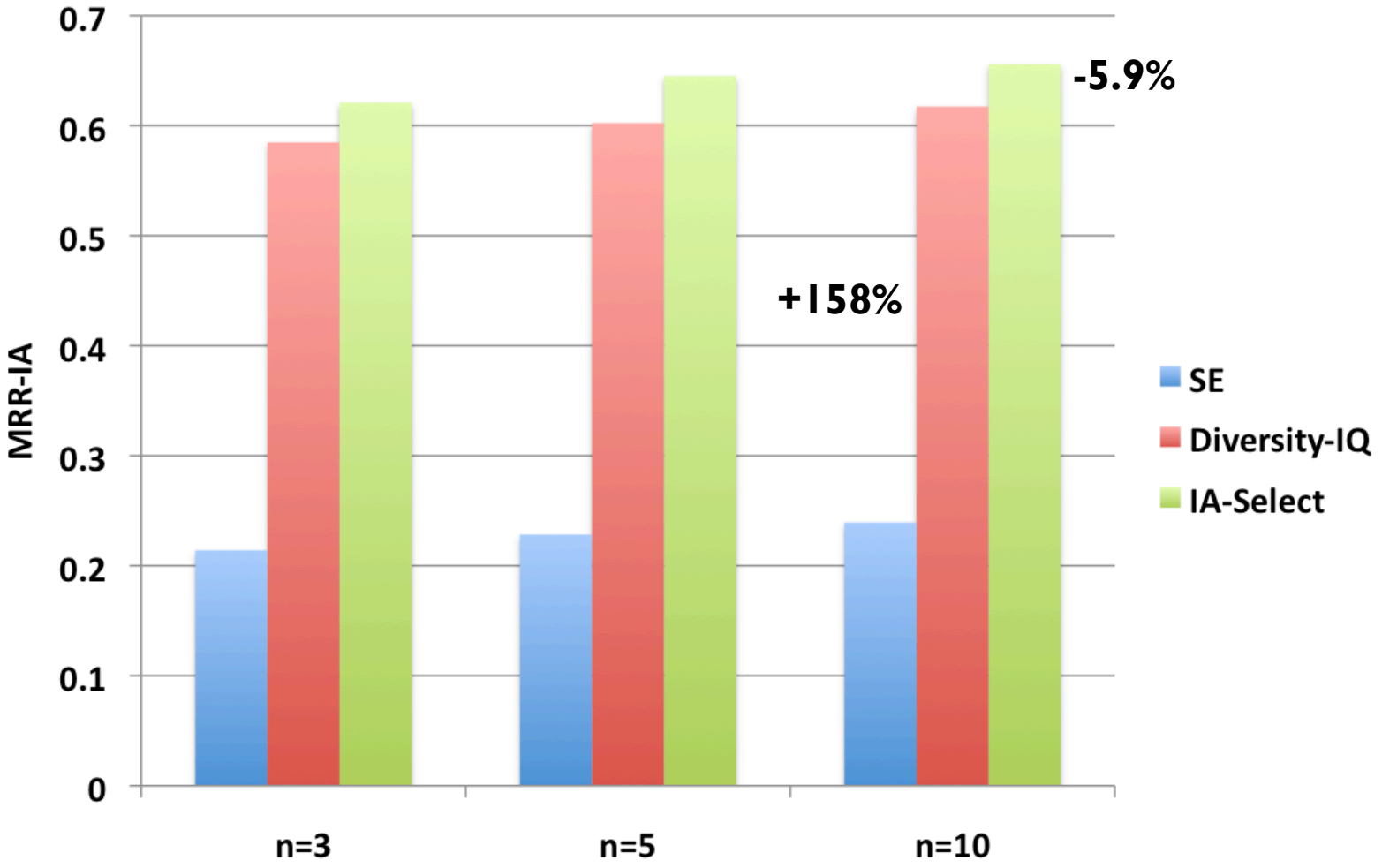
Evaluation Observations

- ***Diversity-IQ* improves expected hits over *SE* and *IA-Select***
  - More expected clickthroughs
- **Performance improvement increases as users are expected to require additional relevant documents**
  - Improved user experience for informational queries

"Single Document" Metrics

- **Compared with metrics which assume a *single relevant document* is sufficient**
  - ***IA-Select* will outperform *Diversity-IQ***
- **Subtopic Recall [ZCL03]**
  - **Measures how quickly the subtopics are covered**
- **Intent-Aware Mean Reciprocal Rank [AGH09]**
  - **MRR, weighted by probability of user intent**

Intent-Aware Mean Reciprocal Rank



Evaluation Observations (cont.)

- **Not surprisingly, *IA-Select* performs better on "single document" metrics**
  - **A trade-off of modeling for informational queries (explicit need for multiple relevant documents)**
  - **If we set our page requirement distribution to Pr(J=1|U) = 1.0, performance is identical**
- ***Diversity-IQ* still outperforms *SE* on both metrics**

Outline

- Identifying localizable queries
- Search result diversity
- **Generating keywords for video**

Video Search Results

baseball bat Search

Josh Womack's crazy **bat** skills at Long Beach Armada 2009

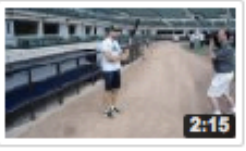
AAA outfielder Josh Womack demonstrates his crazy **bat** skills at Long Beach Armada 2009 Training Camp. Womack's ability to swing the **bat** around and ...

by longbeacharmada | 1 year ago | **3,927,699 views**

HOW A **BASEBALL BAT** IS MADE

Did you ever wonder how a **bat** is made?

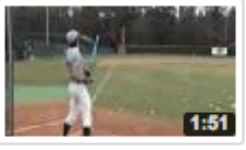
by TheBackstopDOTnet | 2 years ago | **48,836 views**



Josh Womack does a few other **bat** tricks before a Long Beach

Outfielder Josh Womack opens up his bag of tricks for the TV cameras in Edmonton to display his out-of-this-world **bat** tricks before a game between ...

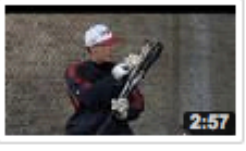
by longbeacharmada | 1 year ago | **844,667 views**



Rawlings Rush **Baseball Bat** Video : BaseballExpress.com

Rawlings Rush **Baseball Bat** Video. high school kids hitting bombs. Brought to you by BaseballExpress.com

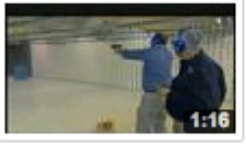
by TeamExpress | 3 years ago | **72,973 views**



Aluminum vs. Wood **Baseball bats**

Former Houston Astro Ron Cacini discusses the differences and dangers of using an aluminum **baseball bat** vs using a wooden **baseball bat**. By Daily ...

by dhphotovid | 2 years ago | **56,163 views**



P226 vs. **Baseball bat**

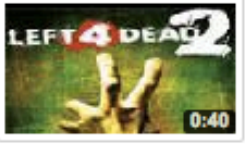
P226 vs. **Baseball bat**

by NallePu | 3 years ago | **87,996 views**

Smashable **Baseball Bat!** HD

I had to reupload...Music problems *NEW CAULK MUG* www.zazzle.com Please Subscribe because it will help me alot! Twitter: www.twitter.com Blog ...

by ParkersTutorials | 1 year ago | **26,164 views**



Left 4 Dead 2: **Baseball Bat** TV Spot Trailer [HQ]

Left 4 Dead 2 TV Spot Trailer [HQ] Developer: Valve Release: 11/17/2009 Genre: FPS Platform: X360/PC Publisher: Valve Website: www.l4d.com/ The ...

Full Episode: Season 1, Episode 25 | by machinima | **Game Trailers** | **117,957 views**

Show other episodes of Game Trailers

Current Limitations

- **Keywords limited to manually entered text**
  - **Title, summary, comments, etc.**
- **Over 10 billion videos watched each month**
  - **Human tagging infeasible at that scale**
- **How do we manually index a full length movie?**
  - **Keywords only relevant over certain segments**
- **Need *automatic* methods for generating keywords from the video *content***
  - **Text content is generally the most accurate**
    - **Scripts, closed captioning tracks, or speech transcripts**

Main Challenges

- **Given the text from a video, how can a search engine identify the meaningful keywords?**
  - **Sources are plaintext, standalone, sparse, and noisy**
- **Vocabulary impedance problem**
  - **Mismatches between content and search keywords**
  - **Can a search engine generate additional, related keywords to improve matching?**

Related Work

- **Keyword identification on Web pages**
  - **HTML features, anchor text, etc. [FPW99, Tur03, KL05, YGC06]**
- **Vocabulary impedance problem**
  - **Augment a page with "neighboring" pages [RCG05]**
  - **Machine translation LM approach [RBG10]**
  - **Term graphs, random walks [LZ01, CC05, JM06, AH07]**
  - **Co-occurrence in retrieved documents [BSA94, SH06]**
  - **Query logs, reformulations [JRM06]**
- **Tag generation using "similar" videos**
  - **Augment keywords from STT [MMH08]**
  - **Apply tags from neighboring pages [SSS09]**

Example Video Text Content

Closed Captioning

00:21:12,897 --> 00:21:14,833
- What do you mean?
- When you think about it,

00:21:14,900 --> 00:21:19,771
- it's as arbitrary as drinkin' coffee.
- Oh. Yeah. Okay.

00:21:19,837 --> 00:21:22,874
Uh, right, then.

Speech Transcript

1271469	40	<SIL>	9
1271510	299	share	33
1271809	49	<SIL>	7
1271859	240	this	27
1272099	340	<SIL>	56
1272439	1280	<s>	39
1273719	310	<SIL>	42
1274030	190	we	44
1274219	199	think	47
1274419	220	about	23
1274640	99	it	82
1274739	190	says	36
1274929	500	Archer	37
1275429	40	<SIL>	29
1275469	359	Street	34
1275829	40	<SIL>	4
1275869	440	coffee	49
1276309	1920	<s>	49

Identifying Relevant Keywords

- **Parse and tag the input text**
  - **Scripts formatted in human readable layout**
    - **Identify scene headings, character names, dialog lines, action descriptions, etc.**
  - **Closed captioning and speech transcripts contain non-text data (time codes, confidence values, etc.)**
- **Construct statistical N-gram tree of length N=4 [CD07]**
  - **Prune tree to select most frequent keywords**

Identifying Keywords From Noisy Data

- **N-gram method requires sufficient amount of (reasonably accurate) input text**
  - **User generated content is often short (3-4 mins)**
  - **Speech transcripts are noisy**
- **Generative method based on topic modeling**
  - **Assumes text is generated by sampling from a few hidden topics (represented as keyword probabilities)**
  - **Identify these topics to help determine relevant keywords from the noise**

Mining for Related Terms

- **Vocabulary impedance problem**
  - **Keywords chosen by authors, actors are not always the same as those chosen by searchers, advertisers**
- **Given keywords from the source text, how can the search engine identify additional relevant keywords?**
- **Consider two related term mining approaches**
  - **Using Web search results**
  - **Using Wikipedia graph structure**

Mining From Web Search Results

- **Semantically similar queries will produce textually similar documents [SH06]**
- **Submit term *T* as a search query**
  - **Frequently co-occurring terms on the result pages are likely to be related**
- **From each result page, identify top keywords**
- **Compute a score for each keyword**
  - **For our evaluations, score is based on corpus frequency (CF) and inverse document frequency (IDF)**

Mining From Wikipedia

- **Graph structures can indicate relationship between terms**
  - **Model Wikipedia as directed graph $G = \{V,E\}$**
- **Identify node t for term T**
  - **Using the page title**
- **Identify nodes forming a direct cycle with t**
  - **(n,t) and (t,n) are both in E**
- **Rank terms $\{n\}$ according to their PageRank**

Merging Multiple Ranked Lists

- **Related keywords from each method are scored on different scales**
  - **CF*IDF vs. PageRank**
- **Only commonality is their relative ranking**
  - **Assign score to term in list l using its reciprocal rank**

$$s_l(t_i) = \frac{1}{1 + \log i}$$

- **Compute score for each term across all *n* lists**

$$S(t) = \sum_{j=1}^{n} \alpha_j s_j(t)$$

Evaluation Setup

- **Evaluated keywords generated for 20 videos with user survey**
  - **Shown full clip or film trailer (3-4 minutes)**
  - **Displayed 5 of top 20 keywords from both keyword generation methods for each text source**
  - **Displayed 1 of the top 10 related terms for each source keyword**
  - **23+ participants from UCLA CSD, social networks**
  - **Minimum of 9 and average of 13 persons evaluating each video**

Evaluation Metrics for Relevancy

- **Relevancy of the keywords**

$$\text{Precision}(S) = \frac{1}{i} \sum_{i} \frac{|K_i(S) \cap R_i|}{|K_i(S)|}$$

$$\text{Potential}(S) = \frac{|R(S)|}{|K(S)|}$$

- **$K_i(S)$ - keywords shown in evaluation i**
- **R_i - keywords marked relevant in evaluation i**
- **K(S) - keywords displayed at least once**
- **R(S) - keywords judged relevant at least once**

Relevancy of Keywords from Source Text

Precision

	Statistical	Generative
Script	0.389	0.353
Closed captioning	**0.443**	0.397
Speech transcript	0.291	0.307

Potential

	Statistical	Generative
Script	0.662	0.635
Closed captioning	0.758	0.705
Speech transcript	0.467	0.514

Relevancy for Speech Transcripts

Precision

	Statistical	Generative
Studio films	0.268	0.252
News and educational	0.442	0.473
User generated	0.268	**0.368**

Relative precision (vs. closed captioning)

	WER	Statistical	Generative
Studio films	0.857	0.723	0.690
News and educational	0.406	0.731	0.961

Relevancy of Related Keywords

Precision

	Statistical-Related	Generative-Related
Script	**0.254**	0.215
Closed captioning	**0.260**	0.221
Speech transcript	0.208	0.186

Observations on Relevance Metrics

- **Statistical N-gram method better for long or well formed text**
- **Generative method appears to be a better choice for noisier data (e.g. speech transcripts)**
- **Relative performance of STT vs. CC is promising, even with high word error rates**
  - **Nearly identical precision for news videos**
- **Related keywords have lower precision**
  - **Might not be accurate enough for search**

Evaluation Metrics for Advertising

- **Usefulness of the keywords to advertisers**

$$\text{Appeal}(S) = \frac{|R(S) \cap A^*|}{|R(S)|}$$

$$\text{Popularity}(S) = \frac{1}{|R(S)|} \sum_{k \in R(S)} A_k^*$$

- **A^* - all keywords which return at least one ad**
- **A_k^* - number of ads returned by keyword *k***
  - **Search engine shows a maximum of 8 ads per query**

Advertising Utility of Keywords

Appeal

	Statistical	S-Related	Generative	G-Related
Script	0.726	0.788	0.607	0.792
Closed captioning	0.578	**0.785**	0.543	**0.796**
Speech transcript	0.681	**0.827**	0.594	**0.820**

Popularity

	Statistical	S-Related	Generative	G-Related
Script	3.59	3.96	3.00	**4.18**
Closed captioning	2.11	**3.81**	2.00	**3.77**
Speech transcript	2.54	**4.39**	2.56	**4.30**

Popularity for Speech Transcripts

	Statistical	S-Related	Generative	G-Related
Studio films	2.97	**4.35**	2.67	**4.39**
News and educational	1.69	**4.11**	2.21	**3.50**
User generated	1.89	**4.83**	2.63	**4.75**

Precision vs. Popularity

- **Trade-off between precision and popularity**
- **Precision-weighted popularity measures the average popularity of the keywords, weighed by their individual precision**

$$PWP(S) = \frac{\sum_{k \in K(S)} A_k^* \; P(S,k)}{|K(S)|}$$

$$P(S,k) = \frac{\sum_{i, k \in K_i(S)} |\{k\} \cap R_i(S)|}{|K_i(S)|}$$

Precision-weighted Popularity

PWP by source

	Statistical	S-Related
Script	**1.358**	0.908
Closed captioning	0.964	0.955
Speech transcript	0.661	0.842

PWP for speech transcripts

	Statistical	S-Related
Studio films	0.726	0.663
News and educational	0.546	**1.278**
User generated	0.563	**1.164**

Observations on Advertising Metrics

- **Related keywords appear more meaningful to advertisers**
- **The most precise sources are also the lowest performing for advertising**
  - **Closed captioning and news videos**
- **Related term mining appears most beneficial for speech transcripts**
  - **Particularly for choosing advertising keywords from news or user generated content**

Summary of Contributions

- Proposed framework for identifying implicitly geo-localizable queries
- Helps search engine know when to apply location context to improve result and advertisement relevance
  - Affects 16% of queries
- Up to 94% accuracy in our evaluations

Summary of Contributions (cont.)

- **Presented algorithm for diversifying search results for ambiguous queries**
  - **Approximately 16% of queries are ambiguous**
- **First model which accounts for requirements of informational queries**
- **Up to 50% improvement over modern algorithm**

Summary of Contributions (cont.)

- **Studied keyword selection methods for sparse text content from videos**
- **Helps search engine more effectively index video content and match relevant ads**
  - **Billions of videos watched every day**
- **Demonstrated vocabulary mismatch problems**
  - **Highlighted where related term mining can be most beneficial**

References

- M. Welch and J. Cho. **Automatically Identifying Localizable Queries.** *SIGIR, 2008.*
- M. Welch, J. Cho, and W. Chang. **Generating Advertising Keywords from Video Content.** *CIKM, 2010.*
- M. Welch, J. Cho, and C. Olston. **Search Result Diversity for Informational Queries.** *Submitted to WSDM, 2011.*
- **See references section of dissertation for complete list**

Thank you!