Microsoft Research

Analyzing Linguistic Structure of Web Search Queries

Rishiraj Saha Roy

Ph.D. Student
Computer Science and Engineering
IIT Kharagpur, India

Under the guidance of

Prof. Niloy Ganguly (IIT Kharagpur, India)
and
Dr. Monojit Choudhury (Microsoft Research India)

Problem Statement: A New Language?

- Structure

  samsung galaxy s4 how to config 3g

  - Syntax
    - More than bags-of-words
    - MWEs, dependencies, relative word orders important
    - Less than full NL sentences (fragments only)
    - Less function words than NL
  - Semantics: Relevance of retrieved documents

Problem Statement: A New Language?

- Function
  - Transmission of information
  - Asymmetric communication and heterogeneous agents
- Dynamics
  - Evolving over time with 2-way interaction
  - Users indirectly influencing each other's search patterns through clicks and query completions

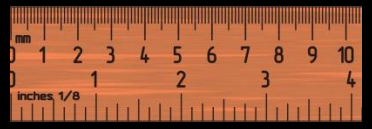
R. Saha Roy, M. Choudhury and K. Bali, "Are Web Search Queries an Evolving Protolanguage?", *in Proceedings of the 9th International Conference on the Evolution of Language 2012 (Evolang IX)*, 13 – 16 March 2012, Kyoto, Japan, pages 304-311.

Motivation

- Query understanding gaining importance [QRU10, QRU11]
  - Vital for solving complex tail queries
- Understanding queries as a language can add a paradigm to the keyword match based retrieval setting where generic language processing tools can be applied (both to queries and documents)
- Sub-problems can prove directly useful for improving IR
- Interesting to study how millions of users, without direct interaction, are showing structural convergence
  - Perfectly preserved dataset for studying language evolution

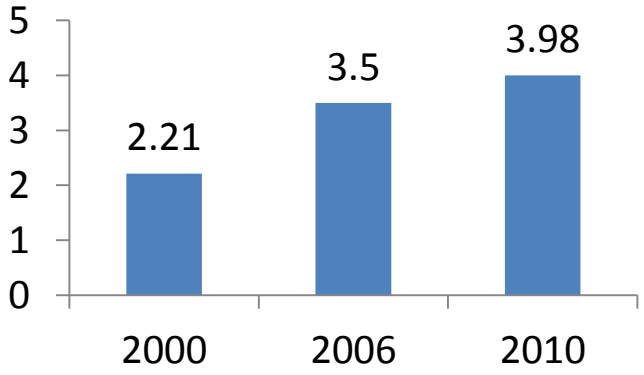
Datasets

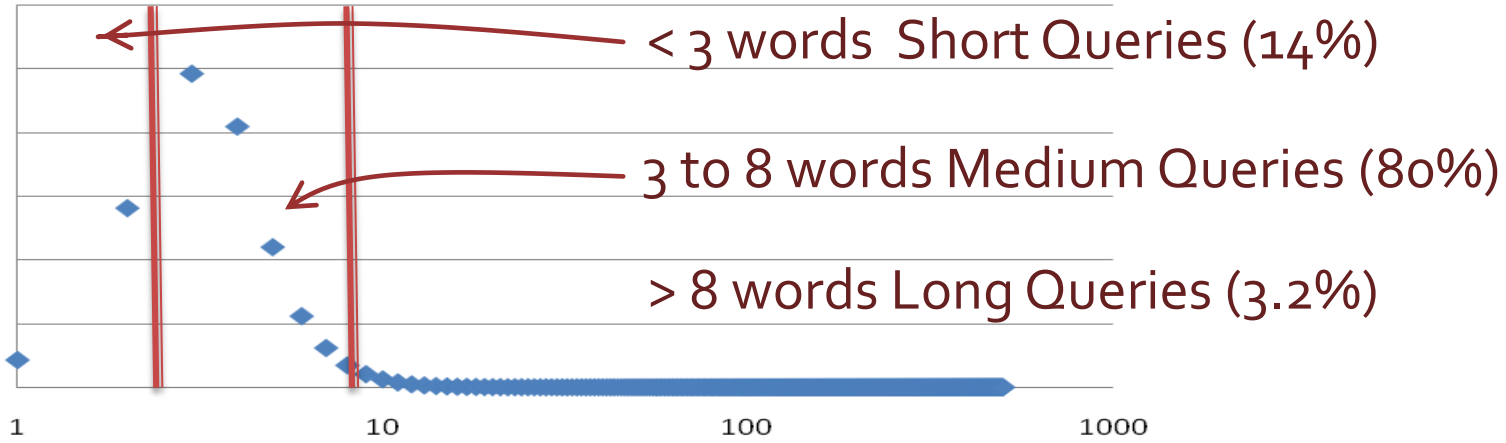
- **Dataset 1:** 340 million query-URL pairs from Bing Australia with relevance scores (March 2010)
- **Dataset 2:** 16.7 million queries, URLs, and other features from Bing Australia (May 2010)
- **Dataset 3:** Summary of searches from 18 domains (Bing Australia, 2010)
- **Thanks to:** Microsoft India Development Center, Hyderabad, India



Query Lengths

The mean length of Web search queries is increasing [Spinko1, Passo6, own data]





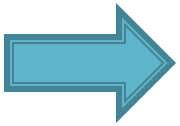
Proposed Approach

- Apply well-established principles from natural language understanding to Web search queries
- ***Unsupervised*** segmentation to identify basic structural units
- Unsupervised induction of lexical categories
- Unsupervised (dependency) grammar induction
- Complex network analysis of query logs
- Understanding query generation

Unsupervised Query Segmentation Using Query Logs

Unsupervised Query Segmentation

2 large islands in the atlantic
i need a crack for windows 7
5 bedroom accommodation hobart large family
cant view videos on youtube
finding a word in text files
american history in 1880 large disappearances
increasing usable ram on windows 7
convert text files to pdf format
movie from rapidshare cant view
improvements in windows 7 official release
3gs mobile phone bluetooth large screen
dead space text files walkthrough ps3
advantages of living in large cities
cant view high resolution videos
import multiple text files into access
import video from camcorder windows 7
facebook photos cant view
index dat files in windows 7
why cant view friends profiles
how to compare 2 text files



cant view | *large* | *text files* | *windows 7*

Discovering Segments

- **Basis:** If "*leonardo da vinci*" is a segment (phrase that cannot be permuted), then a query having *leonardo* and *da* and *vinci* will most likely contain *leonardo da vinci* together

  leonardo da vinci oil paintings

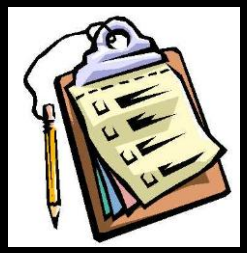
  leonardo di caprio in da vinci code

- Differs from past approaches of MWE (multiword expression) detection [Hagen11, Li11] – does not depend upon frequencies of constituent unigrams
- Interesting: *spot a fake, how to make a, very hungry caterpillar*
- Uses only query logs as the input resource

Discovering Segments (contd.)

- Mathematical Formulation:
  - Compute the expected probability of observing *leonardo da vinci* in queries that contain all the words
  - Compute the observed probability from query log
  - Is observed probability >> expected probability?
  - Use probability bounds to identify significant MWEs

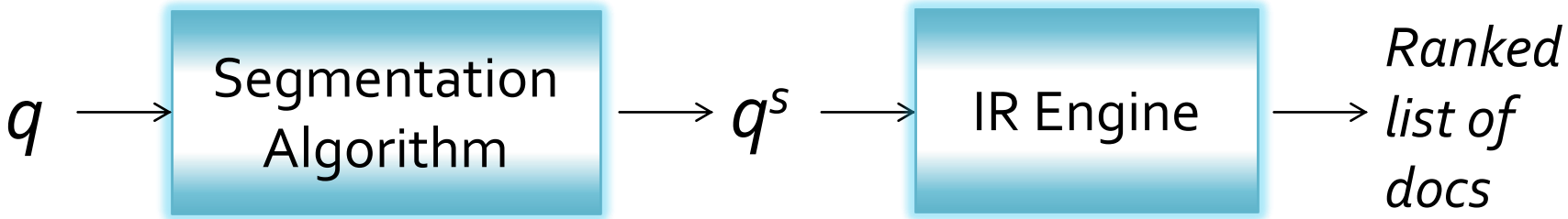
N. Mishra, **R. Saha Roy**, N. Ganguly, S. Laxman and M. Choudhury, "Unsupervised Query Segmentation Using only Query Logs", in Posters of the 20th International World Wide Web Conference 2011 (WWW 2011) Posters, 28 March – 1 April 2011, Hyderabad, India, pages 91 – 92.

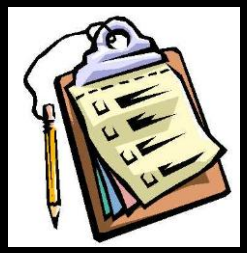


Evaluation of Query Segmentation

- Manually annotated gold standard?
  - High inter-annotator disagreement
  - Not clear what should be the guidelines
  - End-user of segmentation is search engine

An IR-based evaluation scheme



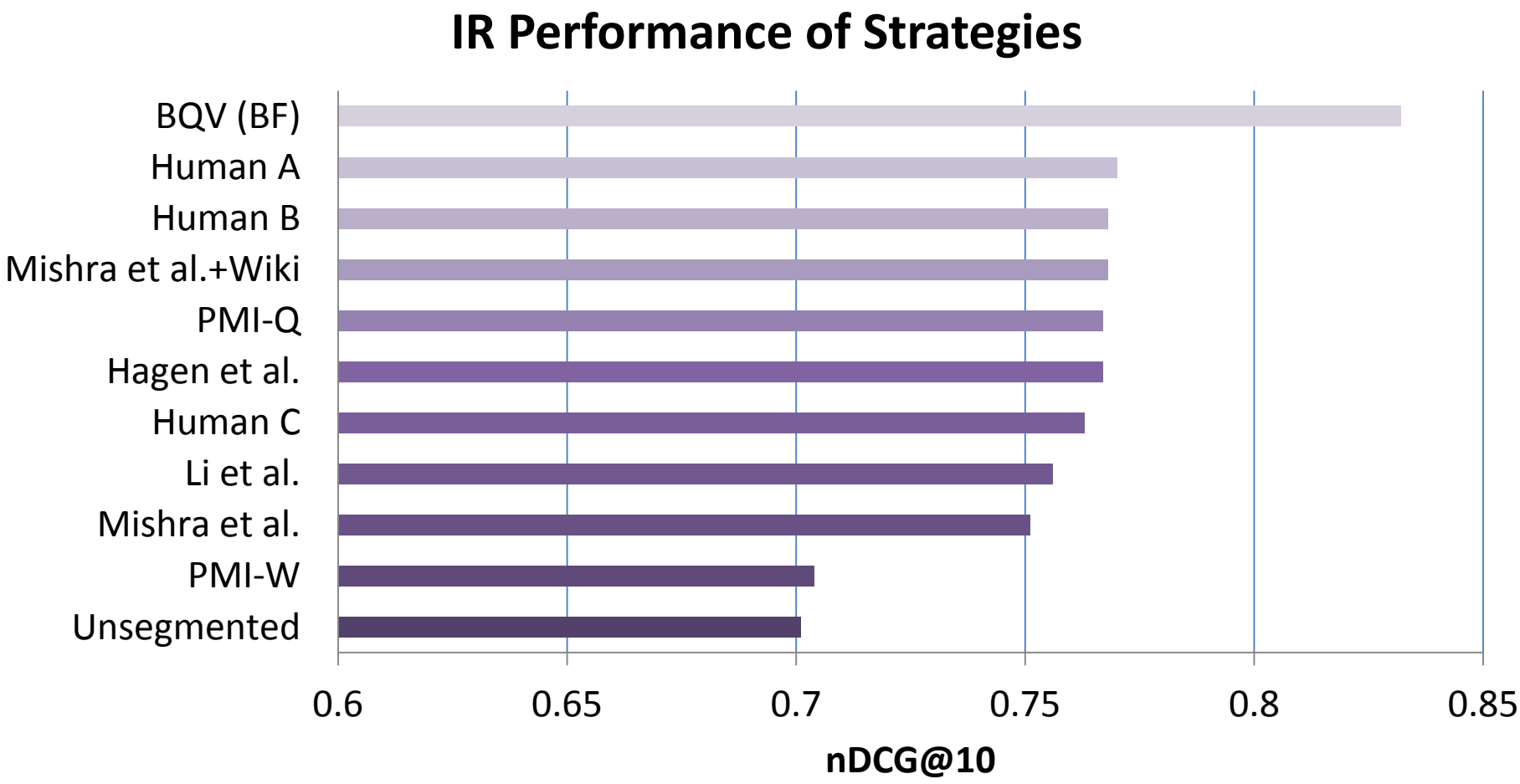


Evaluation of Query Segmentation

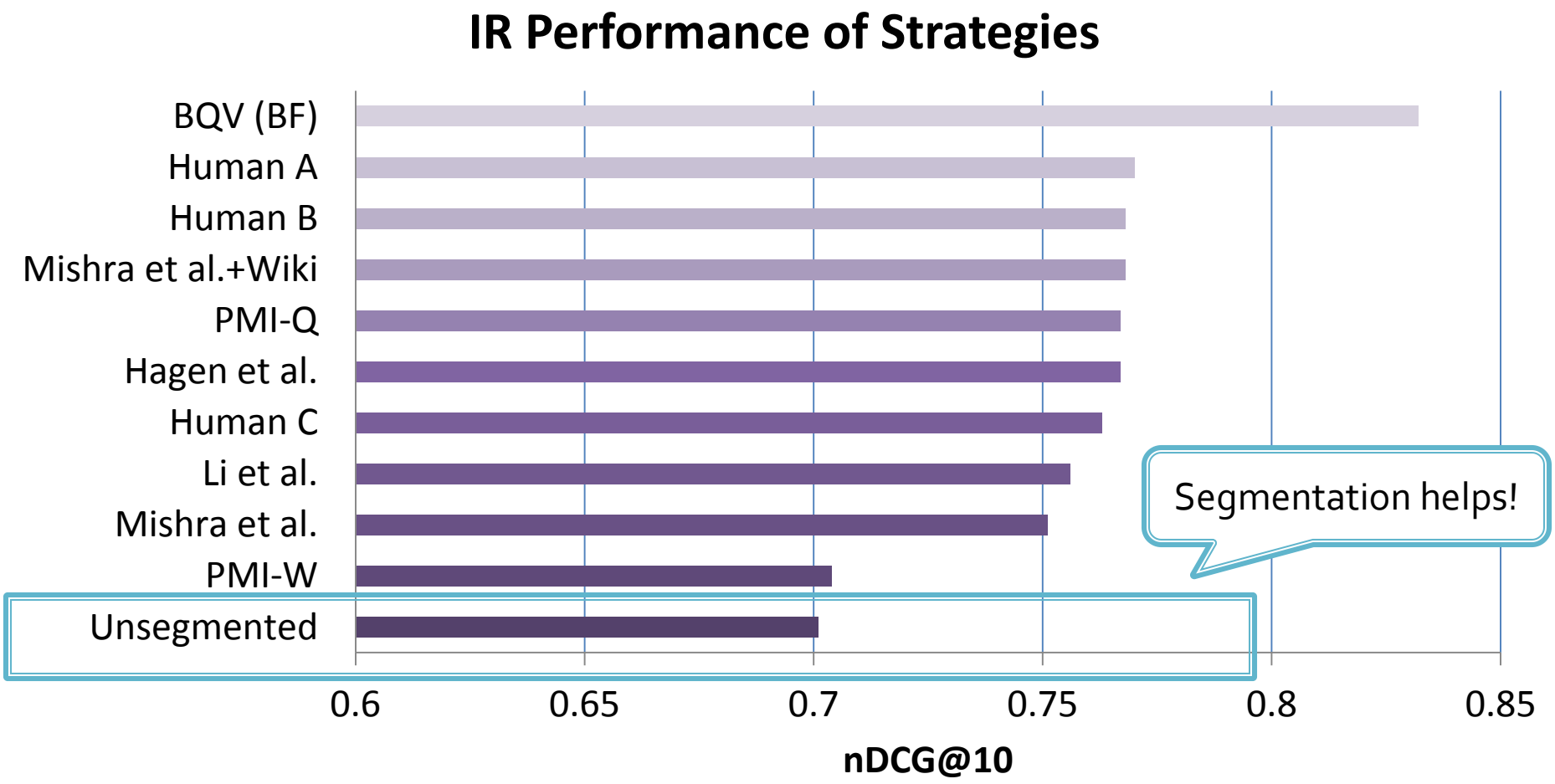
- Use of double quotes
- Generate partially quoted versions
- Evaluate IR performance of each quoted version
- Oracle based approach
- Reflects potential of query segmentation

R. Saha Roy, N. Ganguly, M. Choudhury and S. Laxman, "An IR-based Evaluation Framework for Web Search Query Segmentation", in Proceedings of the 35th Annual ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR '12), Portland, USA, 12 - 16 August 2012, pages 881 – 890.

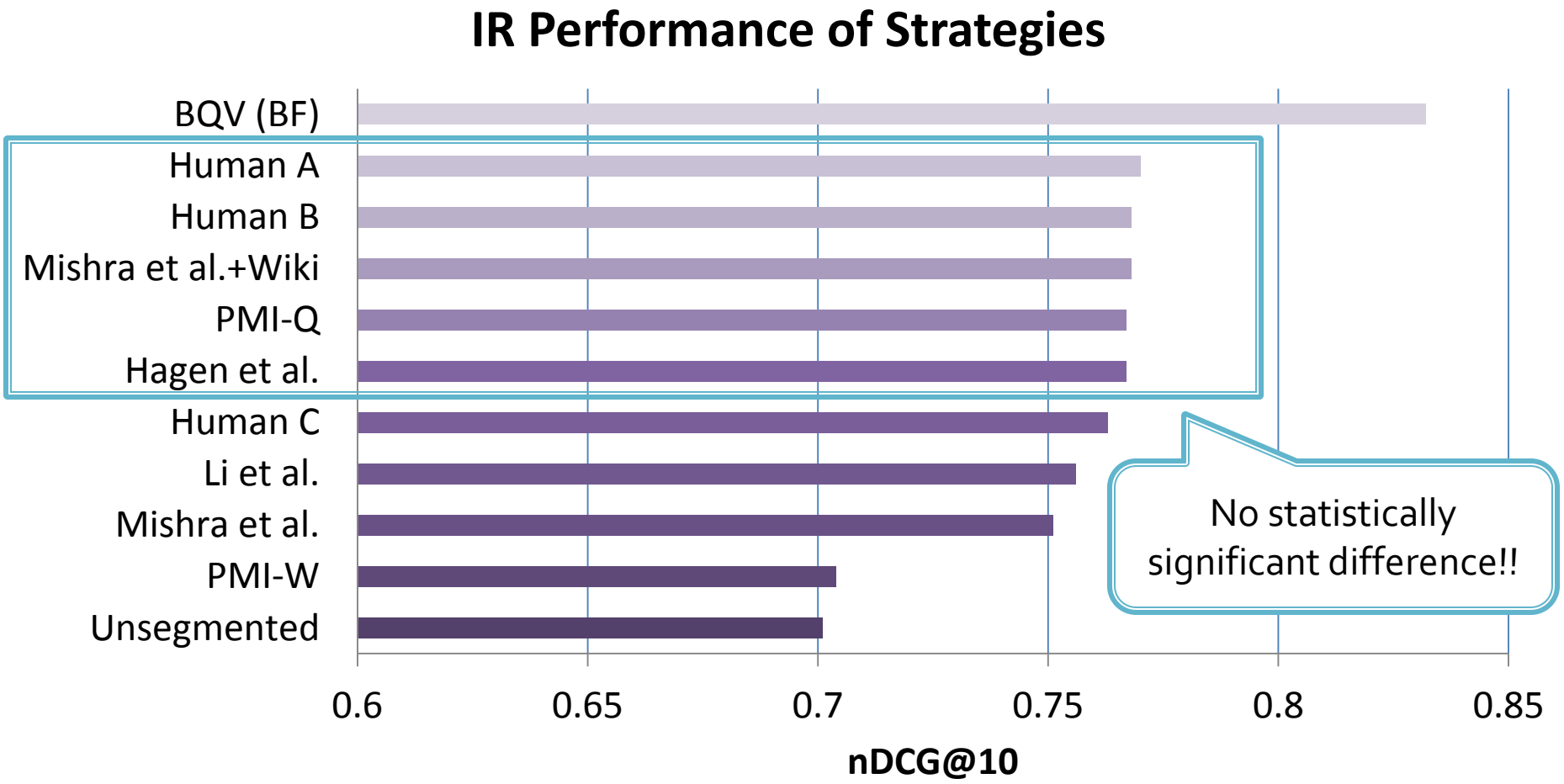
Results


IR Performance of Strategies
BQV (BF)
Human A
Human B
Mishra et al.+Wiki
PMI-Q
Hagen et al.
Human C
Li et al.
Mishra et al.
PMI-W
Unsegmented
0.6
0.65
0.7
0.75
0.8
0.85
nDCG@10

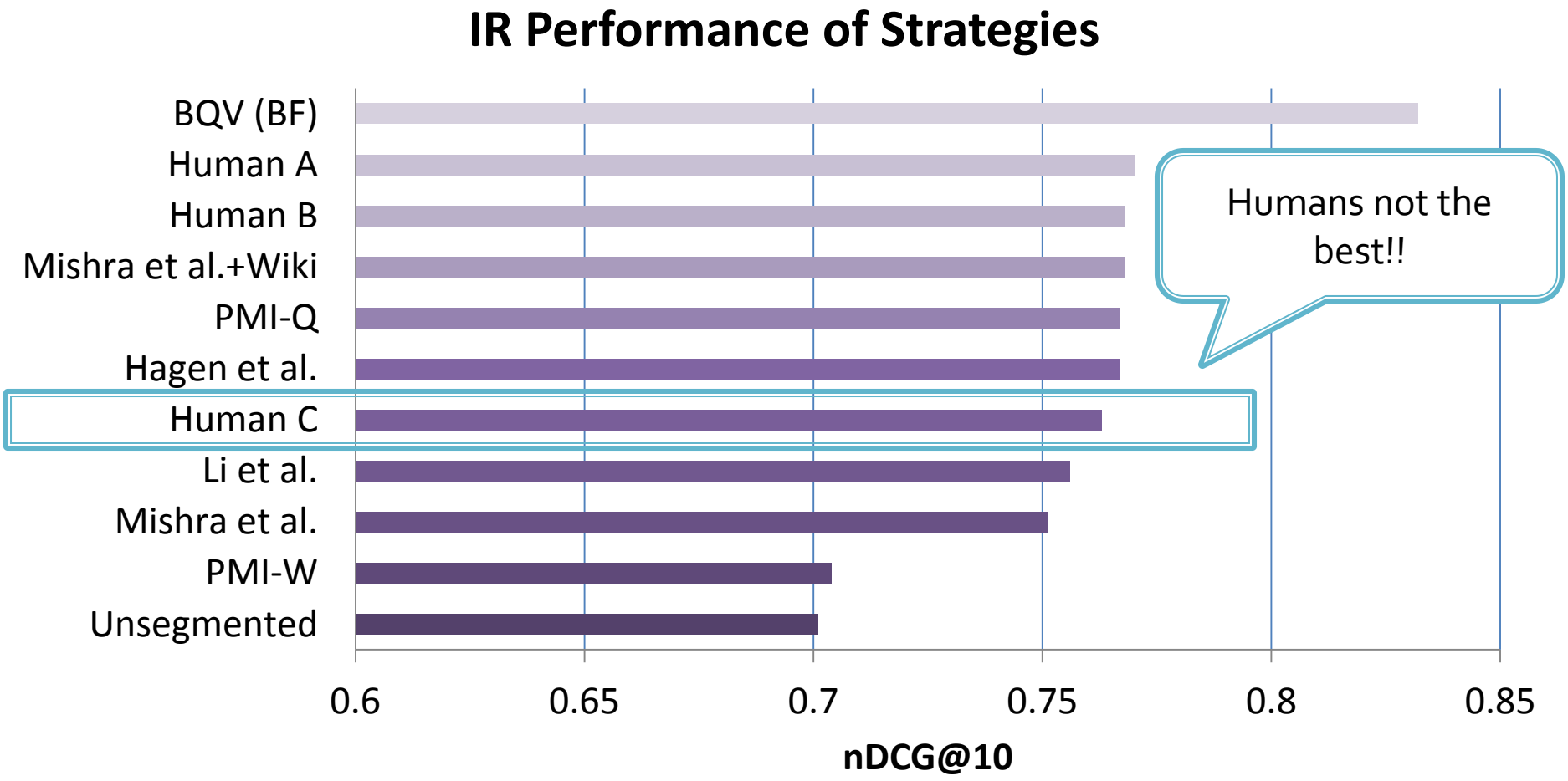
Results


IR Performance of Strategies
BQV (BF)
Human A
Human B
Mishra et al.+Wiki
PMI-Q
Hagen et al.
Human C
Li et al.
Mishra et al.
PMI-W
Unsegmented
Segmentation helps!
0.6
0.65
0.7
0.75
0.8
0.85
nDCG@10

Results


IR Performance of Strategies
BQV (BF)
Human A
Human B
Mishra et al.+Wiki
PMI-Q
Hagen et al.
Human C
Li et al.
Mishra et al.
PMI-W
Unsegmented
No statistically significant difference!!
0.6
0.65
0.7
0.75
0.8
0.85
nDCG@10

Results

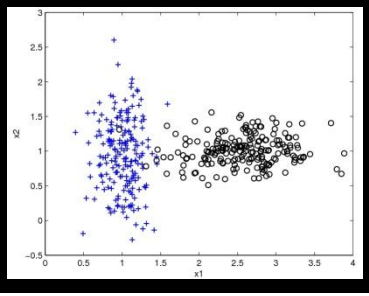

IR Performance of Strategies
BQV (BF)
Human A
Human B
Mishra et al.+Wiki
PMI-Q
Hagen et al.
Human C
Li et al.
Mishra et al.
PMI-W
Unsegmented
Humans not the best!!
0.6
0.65
0.7
0.75
0.8
0.85
nDCG@10

Applications of Query Segmentation

- Improvement in IR precision
- Query expansion
- Can help user in query reformulation
- First step towards query understanding
- NEXT STEP: Nested query segmentation for resolving granularity issues

(samsung (galaxy s4)) ((how to) (config 3g))

Unsupervised Induction of Lexical Categories

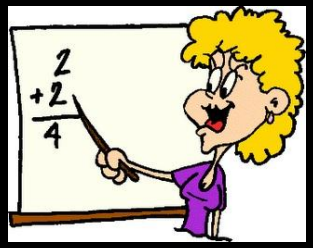


Lexical Categories: Two Kinds of Units

- *how to build a snowman*
- *aborigine history wikipedia*
- *latest news australia bushfires*
- *my heart will go on lyrics*
- *compare canon eos450d with nikon d5000*
- *harry potter and the deathly hallows*

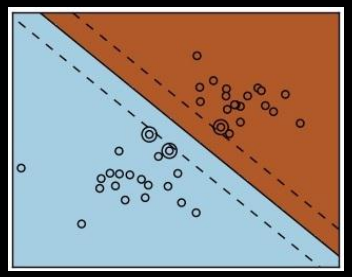
Content units

Intent units



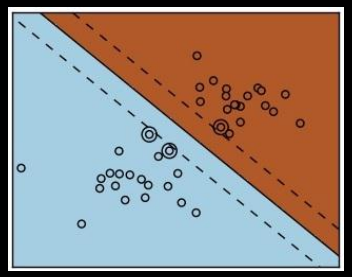
Postulates

- English POS not meaningful for queries [Barr08]
- Queries contain two kinds of units [Yu and Ren12]
- Query semantics dependent on interaction between them
- Content Units
  - Non-negotiable; central concepts in query
  - Present in every relevant document
- Intent Units
  - Not necessarily present in the document
  - Tells us what is required regarding the content units



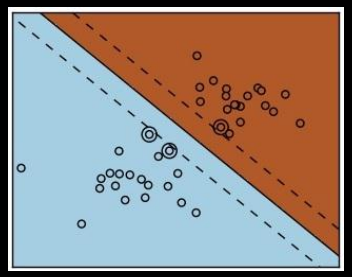
Separating Intent from Content

- Use ***distributional features*** computed from corpus
- Most frequent words expected to be intent units
- More robust statistics: Use co-occurrence properties
  - *nikon camera prices, winter coats prices, property prices in bengaluru, microsoft share prices*
  - *nikon camera prices, nikon camera models, nikon camera for sale, buy nikon camera online*

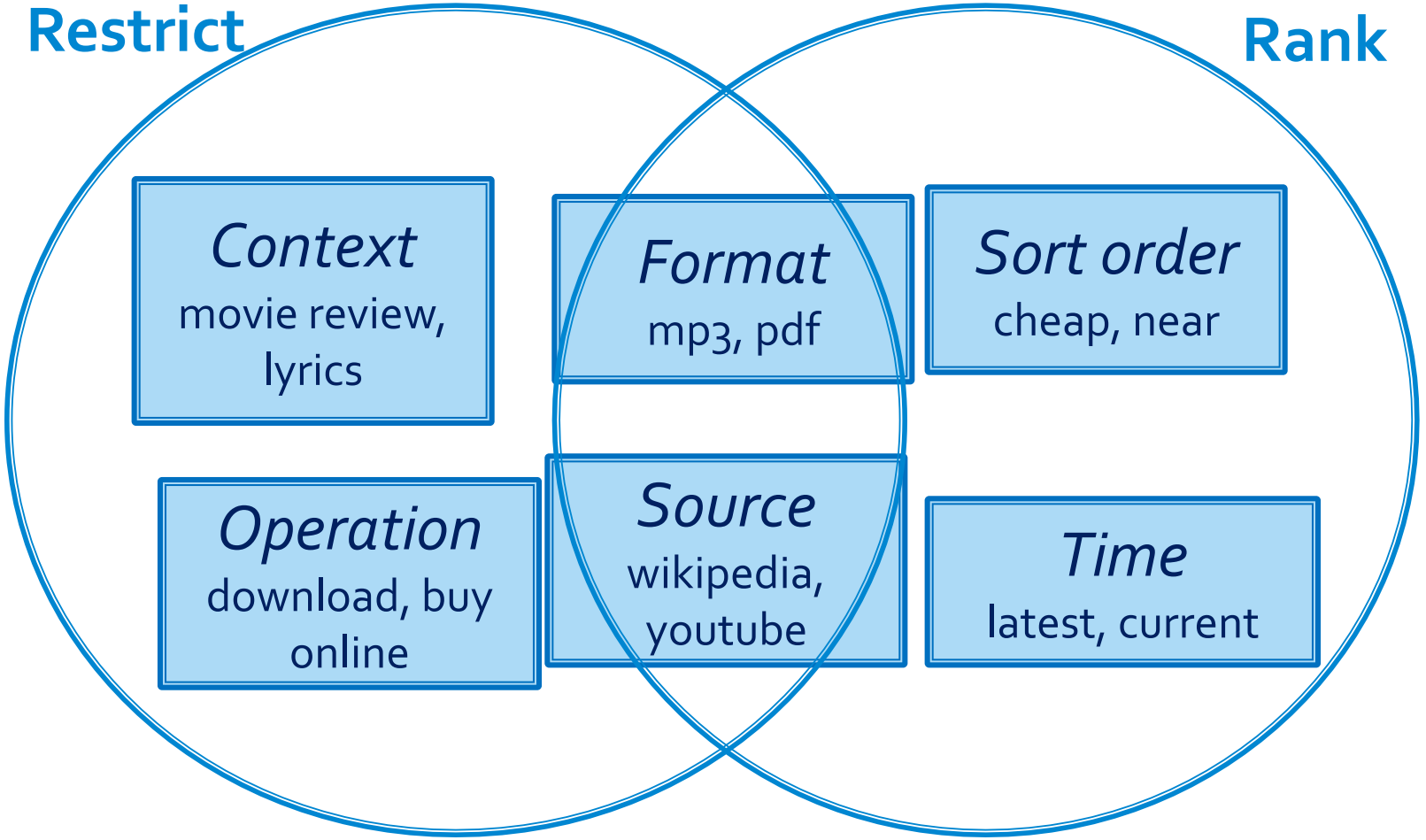


Separating Intent from Content (contd.)

- Co-occurrence counts
- Co-occurrence entropies
- Clustering coefficients (graph-based metric)
- Unsupervised labeling of units in a query with good precision and recall (approximately 70%) against human labeling
- Alternate validation using clickthrough data
  - Intent units have low URL overlap in clickthrough patterns



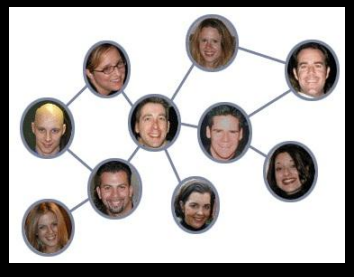
Taxonomy of Intent Units



Labeling of Units is Important

- Applications of detecting intent units
  - Intent units can be processed intelligently to improve relevance
  - Tells us which units to match in the documents
  - Intent diversification
  - Query suggestions
- NEXT STEP: Formulation of simple dependency based ***grammars*** for queries based on content and intent units

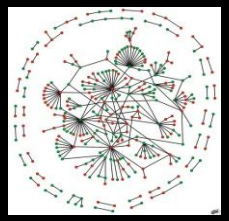
Complex Network Analysis



Complex Networks

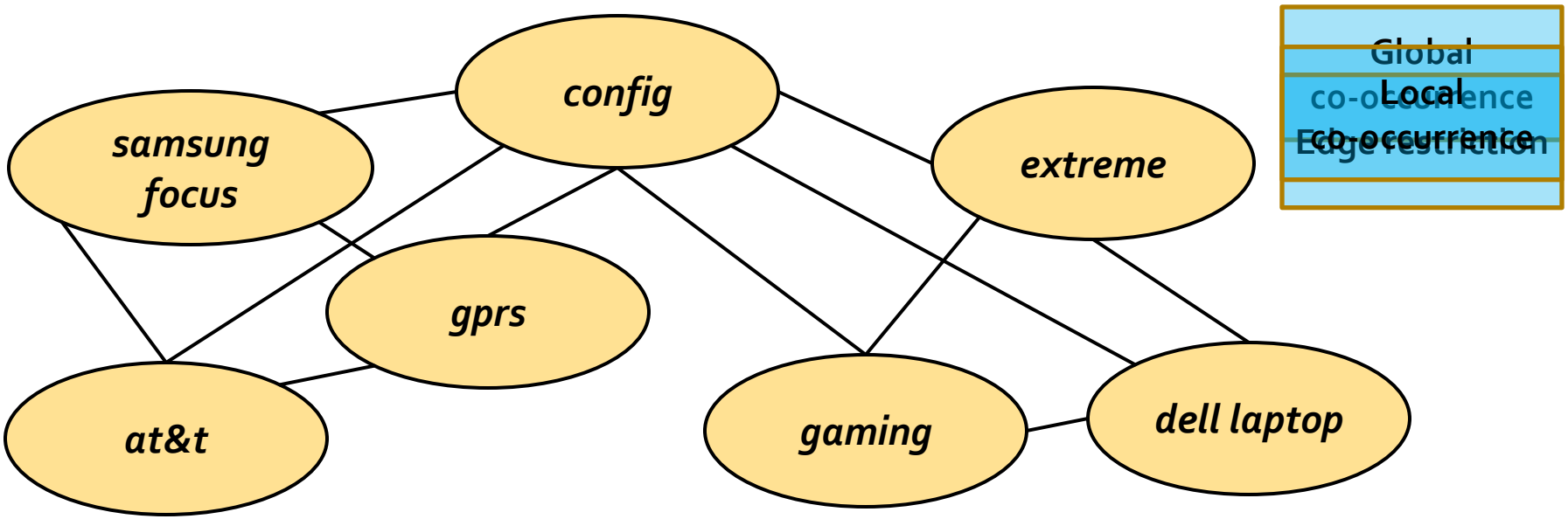
- Real life networks not easily explained by standard topologies
- Applications to linguistics – word co-occurrence networks [Cancho01]
- Interesting tool to discover fundamental properties of a language
- Reveals important corpus-level (global) characteristics
- What would it reveal for queries?

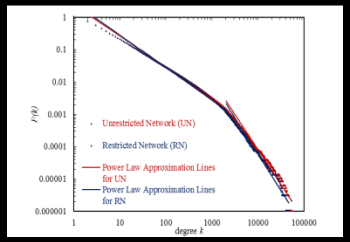
R. Saha Roy, N. Ganguly, M. Choudhury and N. K. Singh, "Complex Network Analysis Reveals Kernel-Periphery Structure in Web Search Queries", in Proceedings of the 2nd International ACM SIGIR Workshop on Query Representation and Understanding 2011 (QRU 2011), 28 July 2011, Beijing, China, pages 5 – 8.



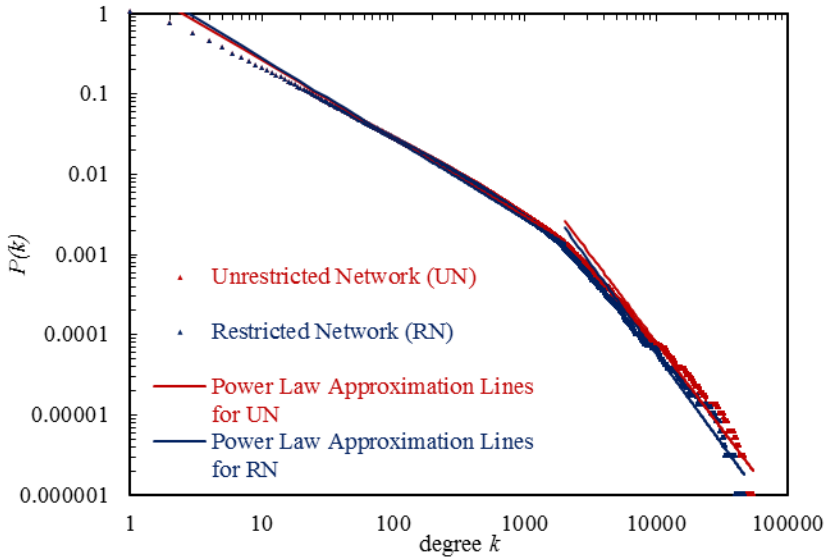
Network Models for Queries

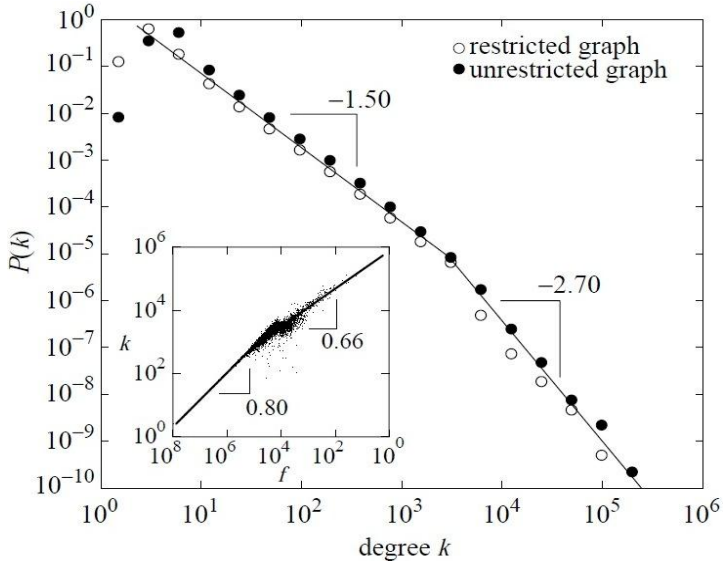
- *"gprs" "config" "samsung focus" "at&t"*
- *"dell laptop" "extreme" "gaming" "config"*



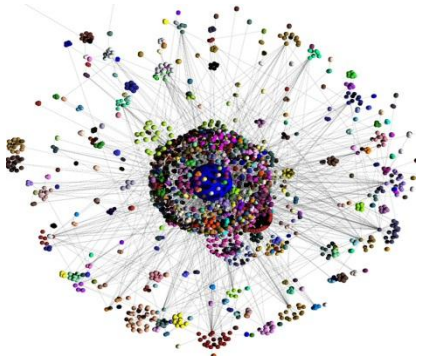


Two-regime Power Law





- Two-regime power law in degree distribution
- Similar coefficients for queries and English
- Kernel (K-Lex) and peripheral (P-Lex) lexicon distinction

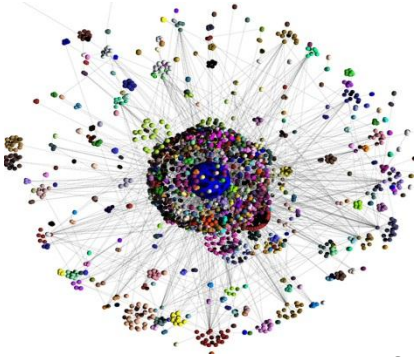


Insights (1)

- ✓ K-Lex and P-Lex
- ✓ Higher mean shortest paths
- ✓ Less tight kernel
- ✓ More k-p edges
- ✓ Socio-cultural effects

- Differences in compositions of K-Lex and P-Lex
- Content and intent units

K-Lex (popular segments)	P-Lex (rarer segments)
how to	*matthew brodrick*
wiki	*accessories*
free	*police officer*
and	*who is*
in australia	*epson tx800*
videos	*star trek next gen*
real estate	*adams apple*
difference between	*harvard university*
windows xp	*leukemia*



Insights (Summary)

- ✓ K-Lex and P-Lex
- ✓ Higher mean shortest paths
- ✓ Less tight kernel
- ✓ More k-p edges
- ✓ Socio-cultural effects

Natural language	Queries
Kernel and periphery	Kernel and periphery
Content and function units both in kernel and periphery	Content and intent units both in kernel and periphery
Kernel – 1000 units	Kernel – 500 units
Periphery – 84000 units	Periphery – 1200000 units
Sentences formed by units from kernel and periphery, or only kernel	Queries mostly formed by units from kernel and periphery, or only periphery
Intra-kernel edges dominate	Kernel-periphery edges dominate
Kernel more tightly coupled	Kernel less tightly coupled

Understanding Query Generation

Generative Models

- What is a good generative language model for queries? [Zhai01]
- Challenge: No generic evaluation framework
- **Complex networks** for macro-level evaluation
- **User studies** (through crowdsourcing) for micro-level evaluation
- Preliminary results show *n*-grams and extensions inadequate
- Modeling constraints based on **interaction of content and intent units** key to better models
- Can help in improving retrieval, query suggestions, query segmentation

Conclusions

- Unsupervised technique of query segmentation using query logs
- IR-based evaluation framework for query segmentation
- Proposed segmentation algorithm outperforms state-of-the-art
- Unsupervised detection and labeling of content and function units
- Segmentation and labeling can improve search performance
- Interesting similarities and differences between queries and NLs using complex networks
- Gained important insights on linguistic structure of queries, at micro- and macro-levels

Future plans

- Develop algorithms for **nested segmentation** and its application to improve retrieval
- Formulate **dependency grammars** for queries based on content and intent segments
- Formulate better **generative models** for queries
- Conduct **cognitive experiments** for understanding human mental model of query formulation

References

- [Barr08] Barr, C., Jones, R., & Regelson, M. (2008, October). The linguistic structure of English web-search queries. In Proceedings of the conference on empirical methods in natural language processing (pp. 1021-1030). Association for Computational Linguistics.
- [Cancho01] i Cancho, R. F., & Solé, R. V. (2001). The small world of human language. Proceedings of the Royal Society of London. Series B: Biological Sciences, 268(1482), 2261-2265.
- [Hagen11] Hagen, M., Potthast, M., Stein, B., & Bräutigam, C. (2011, March). Query segmentation revisited. In Proceedings of the 20th international conference on World wide web (pp. 97-106). ACM.
- [Li11] Li, Y., Hsu, B. J. P., Zhai, C., & Wang, K. (2011). Unsupervised query segmentation using clickthrough for information retrieval. 34th SIGIR, 285-294.
- [QRU10] Croft, W. B., Bendersky, M., Li, H., & Xu, G. (2010, December). Query representation and understanding workshop. In SIGIR Forum (Vol. 44, No. 2, pp. 48-53).
- [QRU11] Li, H., Xu, G., & Croft, B. (2011). Query Representation and Understanding. In Proceedings of the 2nd Workshop on query Representation and Understanding.
- [Zhai01] Lafferty, J., & Zhai, C. (2001, September). Document language models, query models, and risk minimization for information retrieval. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 111-119). ACM.

Questions??


Thank You!