Unsupervised Approaches to Syntactic Analysis of Web Search Queries

Rishiraj Saha Roy and Niloy Ganguly (IIT Kharagpur)

Monojit Choudhury (Microsoft Research India)

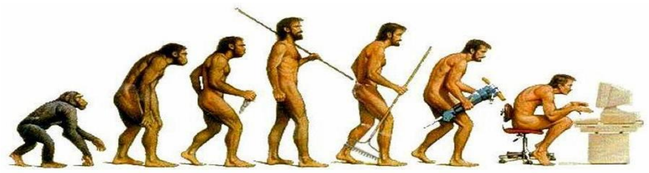


Image source: http://shishikwena.blogspot.in/

Introduction

Web users communicate their information need to a search engine through queries. Queries have a structure far simpler than NL, but more complex than the commonly assumed bag-of-words model. In fact, Web search queries define a new and fast evolving language of its own, whose dynamics is governed by the behavior of the search engine towards the user and that of the user towards the engine.

***Are Web Search Queries* a self-organizing system of language?**

Publications

- Rishiraj Saha Roy, Rahul Katare, Niloy Ganguly, Srivatsan Laxman and Monojit Choudhury, "Discovering and understanding word level user intent in web search queries", in Web Semantics: Science, Services and Agents on the World Wide Web, Elsevier, 2014 (in press).
- Rishiraj Saha Roy, Rahul Katare, Niloy Ganguly and Monojit Choudhury, "Automatic Discovery of Adposition Typology", in Coling '14.
- Rishiraj Saha Roy, M. Dastagiri Reddy, Niloy Ganguly and Monojit Choudhury, "Understanding the Linguistic Structure and Evolution of Web Search Queries", in Evolang X.
- Rishiraj Saha Roy, Yogarshi Vyas, Niloy Ganguly and Monojit Choudhury, "Improving Unsupervised Query Segmentation using Parts-of-Speech Sequence Information", in SIGIR '14 Short Papers.
- Rohan Ramanath, Monojit Choudhury, Kalika Bali and Rishiraj Saha Roy, "Crowd Prefers the Middle Path: A New IAA Metric for Crowdsourcing Reveals Turker Biases in Query Segmentation", in ACL '13.
- Rishiraj Saha Roy, Anusha Suresh, Niloy Ganguly and Monojit Choudhury, "Place Value: Word Position Shifts Vital to Search Dynamics", in WWW '13 Posters.
- Rishiraj Saha Roy, Niloy Ganguly, Monojit Choudhury and Srivatsan Laxman, "An IR-based Evaluation Framework for Web Search Query Segmentation", in SIGIR '12.
- Rishiraj Saha Roy, Monojit Choudhury and Kalika Bali, "Are Web Search Queries an Evolving Protolanguage?", in Evolang IX. [BEST RESEARCH POSTER AWARD].
- Nikita Mishra, Rishiraj Saha Roy, Niloy Ganguly, Srivatsan Laxman and Monojit Choudhury, "Unsupervised Query Segmentation Using only Query Logs", in WWW '11 Posters.

ORIGINAL LOG

larry the lawnmower tv show
our lady of lourdes seven hills
grand theft auto 3 ps2 cheats
lyrics for my name is
never miss a beat mp3 download
room to let perth wa
as time goes by sheet music
villeroy and boch kitchen sinks
we are the people song lyrics
worlds best chocolate chip cookies
another way to die piano sheet
vanilla ice cream in french
piano chords suddenly i see
paris by night sydney 2009

Identifying Syntactic Units

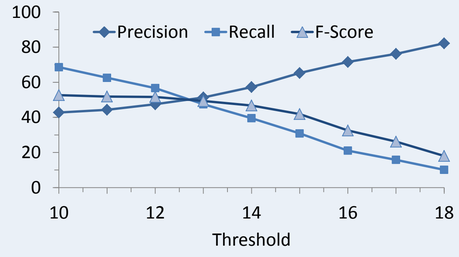
- N: Number of queries in log containing n-gram M
- k: Queries containing words of M in any order
- E: Expected count of M under a bag-of-words *NULL*
- If $\mathbf{P}[E \geq N] < exp(-\frac{2(N-E)^2}{k})$, bounded by a threshold, holds, then M is declared a *query segment*
- Use PMI-based scheme with Wikipedia titles for named entities
- Dynamic programming to search over all possible segmentations

SEGMENTED LOG

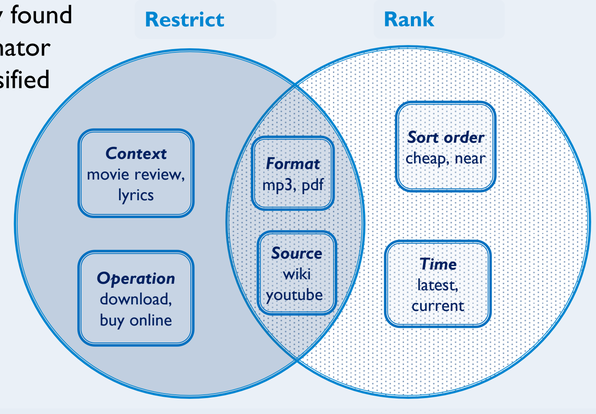
larry the lawnmower | tv show
our lady of lourdes | seven hills
grand theft auto 3 | ps2 cheats
lyrics for | my name is
never miss a beat | mp3 download
room to let | perth wa
as time goes by | sheet music
villeroy and boch | kitchen sinks
we are the people | song lyrics
worlds best | chocolate chip cookies
another way to die | piano sheet
vanilla ice cream | in french
piano chords | suddenly i see
paris by night | sydney 2009

Discovering Syntactic Categories

- Queries contain two types of segments – content and intent
- Analogous to content and function words of NL
- Study distributional statistics of query segments – frequency co-occurrence counts, co-occurrence entropies, and graph clustering coefficients
- Annotate segments manually
- Co-occurrence entropy found to be the best discriminator
- Intent units can be classified



Intent segment detection against human annotations (reverse for content)

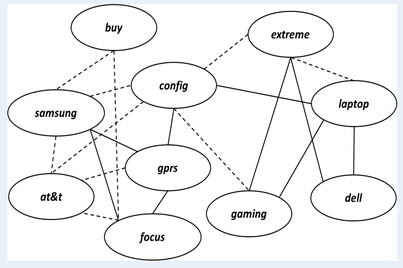


A taxonomy of intent segments in Web search queries.

Applications

- Sponsored search
- Query suggestions
- Intent diversification

Applying Complex Network Modeling



A Word Co-occurrence Network (WCN) built from a toy query log.

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
Small world effect	No small world effect
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

This work has applications in query understanding. We should also use this well-preserved data for studying language evolution.

Contact: rishiraj.saharoy@gmail.com

Microsoft Research