Complex Network Analysis Reveals Kernel-Periphery Structure in Web Search Queries

Rishiraj Saha Roy and Niloy Ganguly
IIT Kharagpur, India

Monojit Choudhury
Microsoft Research India

Naveen Kumar Singh
NIT Durgapur, India

Sparse

Clustering coefficient greater than random

High mean shortest path

Small-world nature

Kernel (popular units) and Peripheral (rarer units) Lexicons

Similar to English; close power law exponents

Two-regime power law

Heads and modifiers

Most edges run between kernel and periphery for queries; intra-kernel edges dominate for English

Complex Networks – Real life networks not easily explained by standard topologies

Use of complex networks has proved to be useful in linguistics

Interaction between search engines and users has led to a new language of queries

Edges may be unrestricted or restricted ($p_{ij} > p_i p_j$)

Results in eight different co-occurrence network models

Co-occurrence relationships may be local (within 2 words) or global (anywhere in query)

Units (words/segments) form nodes; co-occurrence relationships result in edges

The periphery-to-kernel ratio is much higher for queries than for English

All experiments done on 16.7 million Bing queries from Australia

Kernel is much less connected than English

Queries may be formed from peripheral units alone

Inference of context often difficult for queries

Socio-cultural factors govern the kernel-periphery distinction in queries

Query node labels: goldilocks and the three bears, buddhism, programming, pics, wikipedia, in australia, leukemia, near, to, clinic, accessories, cats and dogs, movie, of, airedale terrier, tumor, roger federer, latest, pdf, titanium, blood pressure, download, videos, lord of the rings, how to, high schools, harvard university, angkor wat, world heritage listing, best, is, and, in

Comparison of Topological Characteristics of Word Co-occurrence Networks for Queries and Standard Language

Network	Expansion	C	C_{rand}	d	d_{rand}	k_{cross}	γ_1 (DD)	γ_2 (DD)
$QUWN_l$	Query Unrestricted Word Network (local)	0.57	4.24×10^{-5}	7.08	4.24	1,000	−1.82	−3.11
$QRWN_l$	Query Restricted Word Net work (local)	0.28	3.59×10^{-5}	9.45	4.48	1,000	−1.82	−3.30
$QUWN_g$	Query Unrestricted Word Network (global)	0.61	5.48×10^{-5}	7.00	3.92	1,500	−1.81	−2.89
$QRWN_g$	Query Restricted Word Network (global)	0.48	4.57×10^{-5}	7.12	4.14	1,500	−1.89	−2.85
$QUSN_l$	Query Unrestricted Segment Network (local)	0.44	2.43×10^{-5}	7.19	4.62	1,500	−1.97	−3.13
$QRSN_l$	Query Restricted Segment Network (local)	0.36	2.29×10^{-5}	7.31	4.72	1,500	−2.00	−2.82
$QUSN_g$	Query Unrestricted Segment Network (global)	0.47	2.85×10^{-5}	7.12	4.39	2,000	−1.98	−3.17
$QRSN_g$	Query Restricted Segment Network (global)	0.41	2.71×10^{-5}	7.22	4.46	2,000	−2.01	−3.25
$NUWN_l$	NL Unrestricted Word Network (local)	0.69	1.55×10^{-4}	2.63	3.03	≈ 2,500	−1.50	−2.70
$NRWN_l$	NL Restricted Word Network (local)	0.44	1.55×10^{-4}	2.67	3.06	≈ 2,500	−1.50	−2.70

Degree distribution plot legend: Unrestricted Network (UN); Restricted Network (RN); Power Law Approximation Lines for UN; Power Law Approximation Lines for RN

Contact: Rishiraj Saha Roy (*rishiraj.saharoy@live.in*), Niloy Ganguly (*niloy@cse.iitkgp.ernet.in*), Monojit Choudhury (*monojitc@microsoft.com*)

Microsoft Research