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

Rishiraj Saha Roy and Niloy Ganguly
IIT Kharagpur
Kharagpur, India - 721302.
{rishiraj, niloy}@cse.iitkgp.ernet.in

Monojit Choudhury
Microsoft Research India
Bangalore, India - 560080.
monojitc@microsoft.com

Naveen Kumar Singh
NIT Durgapur
Durgapur, India - 713209.
08-cse-35@nittdgp.ac.in

ABSTRACT

Web search queries have evolved into a language of their own. In this paper, we substantiate this fact through the analysis of complex networks constructed from query logs. Like natural language, a two-regime degree distribution in word or phrase co-occurrence networks of queries reveals the existence of a small kernel and a very large periphery. But unlike natural language, where a large fraction of sentences are formed only using the kernel words, most queries consist of units both from the kernel and the periphery. The long mean shortest path for these networks further show that paths between peripheral units are typically connected through nodes in the kernel, which in turn are connected through multiple hops within the kernel. The extremely large periphery implies that the likelihood of encountering a new word or segment is much higher for queries than in natural language, making the processing of unseen queries much harder than that of unseen sentences.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Query Formulation

General Terms
Measurement, Experimentation

Keywords
Web search queries, Co-occurrence networks, Two-regime power law, Kernel and peripheral lexicons

1. INTRODUCTION

Web users communicate their information need to a search engine through queries. The fact that search engines do not really understand or process natural language grammar drives the average Web user to specify their queries in a language that has a structure far simpler than natural languages, but perhaps 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 are governed by the behavior of the search engine towards the user and that of the user towards the engine. Earlier works have shown that in general, direct application of grammatical frameworks and concepts used in Natural Language Processing to understand queries has not been very productive. The linguistic structure of queries has significant differences and similarities from that of natural languages and has to be discovered and understood from the first principles. Not much linguistic or statistical analysis has been done to understand this structure, mainly because, query log data is not publicly available. Moreover, the absence of a generic query structure across all domains make fully automated techniques for linguistic analysis infeasible.

In recent times, complex networks have provided an elegant framework for understanding and analyzing evolving linguistic structures. For example, modeling syntactic and semantic features of a language, consonant inventories and their dynamics can all be accomplished through the study of linguistic networks [1]. Of special interest are the word co-occurrence networks, where each distinct word is considered a node and edges represent co-occurrence of two words in the same sentence. [2]. Similarly, in our models, each query is viewed as a sentence. We use two notions of co-occurrence – local and global. In the local co-occurrence networks, immediate word neighborhood is considered important and an edge is added between two words only if they occur within a distance of two (i.e. separated by zero or one word) in a query [2]. Global networks assume order independence; edges are added between words if they are found in the same query. Restriction of a network is used to prune the edges which might occur purely by chance [2]. To be precise, let $i$ and $j$ be two distinct words from the corpus. Let $p_i$, $p_j$ and $p_{ij}$ be the probabilities of occurrence of $i$, $j$ and the bigram $ij$ in the data. Then, in a restricted network, an edge exists if $p_{ij} > p_i p_j$.

Several researchers have argued that queries are bags-of-segments, and not bags-of-words [3, 4]. For example, the query "australian open 2011 home page" is semantically equivalent to the queries "australian open 2011" "home page" and "home page" "australian open 2011", but not to "page 2011 open australian home". Therefore, we also study the co-
occurrence networks of segments, where nodes are distinct segments instead of words, with the same principles used for edge creation. Query segmentation has been performed based on a technique that uses query logs as the only resource for training [4]. This method is fundamental to our goal because the use of document-based resources for query segmentation risks projection of natural language structure onto queries. For example, the likely segmentation of a query using document resources would be “how to” “spot” “a fake” “bill”. This is segmented as “how to” “spot a fake” “bill” by our method. While a fake bill is a noun phrase, and therefore, a valid segment according to the Standard English grammar, one cannot deny the fact that how to expresses a class of intent in queries and is found to be associated with diverse concepts such as save money, play guitar or make tea [4]. Thus, we formulate eight different complex network models for queries, as described in Table 1. The last two networks were built analogously from natural language text from 10^7 words of British National Corpus (http://www.natcorp.ox.ac.uk/) [2].

These networks have been constructed out of 16.7 million queries collected through Bing Australia\(^1\) during the period February 1, 2009 to June 6, 2009. We removed queries with non-ASCII characters and those having one or more than ten words, resulting in 11.9 million queries for construction of the networks. Single word queries do not participate in co-occurrence relationships and those longer than ten words (very few in number) are mostly constituted of computer generated error messages.

Let \( |N| \) and \( |E| \) be the number of nodes and edges in the network respectively. \( |N| \) is 528,701 and 771,100 for the word and segment networks respectively. This difference is due to the fact that each sub-sequence of words in a multi-word segment can also form valid segments. \( |E| \) varies between 5 million and 8.5 million for the eight networks. Table 1 presents some basic topological statistics of the eight networks and also the corresponding values for word co-occurrence networks built from natural language corpus (as reported in [2]). Here, \( C \) denotes the average clustering coefficient [5], \( d \) the mean shortest path between nodes, and \( C_{\text{rand}} \) and \( d_{\text{rand}} \) represent the corresponding values for a random graph, where \( C_{\text{rand}} \sim k/|N| \) [2] and \( d_{\text{rand}} \sim \ln |N|/\ln (k) \) [5] (where \( k \) is the average degree of occurrence networks build from natural language sentences

\[^1\]http://www.bing.com/?cc=au

---

### Table 1: Complex network models for query logs and natural language.

<table>
<thead>
<tr>
<th>Network</th>
<th>Expansion</th>
<th>( C )</th>
<th>( C_{\text{rand}} )</th>
<th>( d )</th>
<th>( d_{\text{rand}} )</th>
<th>( k_{\text{cross}} )</th>
<th>( \gamma_1(\text{DD}) )</th>
<th>( \gamma_2(\text{DD}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>QUW (_N_1)</td>
<td>Query Unrestricted Word Net (local)</td>
<td>0.57</td>
<td>4.24 \times 10^{-5}</td>
<td>7.08</td>
<td>4.24</td>
<td>1000</td>
<td>-1.82</td>
<td>-3.11</td>
</tr>
<tr>
<td>QRW (_N_1)</td>
<td>Query Restricted Word Net (local)</td>
<td>0.28</td>
<td>3.59 \times 10^{-5}</td>
<td>9.45</td>
<td>4.48</td>
<td>1000</td>
<td>-1.82</td>
<td>-3.30</td>
</tr>
<tr>
<td>QUW (_N_2)</td>
<td>Query Unrestricted Word Net (global)</td>
<td>0.61</td>
<td>5.48 \times 10^{-5}</td>
<td>7.00</td>
<td>3.92</td>
<td>1500</td>
<td>-1.81</td>
<td>-2.89</td>
</tr>
<tr>
<td>QRW (_N_2)</td>
<td>Query Restricted Word Net (global)</td>
<td>0.48</td>
<td>4.57 \times 10^{-5}</td>
<td>7.12</td>
<td>4.14</td>
<td>1500</td>
<td>-1.89</td>
<td>-2.85</td>
</tr>
<tr>
<td>QUUSN (_N_1)</td>
<td>Query Unrestricted Segment Net (local)</td>
<td>0.44</td>
<td>2.43 \times 10^{-5}</td>
<td>7.19</td>
<td>4.62</td>
<td>1500</td>
<td>-1.97</td>
<td>-3.13</td>
</tr>
<tr>
<td>QRSN (_N_1)</td>
<td>Query Restricted Segment Net (local)</td>
<td>0.36</td>
<td>2.29 \times 10^{-5}</td>
<td>7.31</td>
<td>4.72</td>
<td>1500</td>
<td>-2.00</td>
<td>-2.82</td>
</tr>
<tr>
<td>QUUSN (_N_2)</td>
<td>Query Unrestricted Segment Net (global)</td>
<td>0.47</td>
<td>2.85 \times 10^{-5}</td>
<td>7.12</td>
<td>4.39</td>
<td>2000</td>
<td>-1.98</td>
<td>-3.17</td>
</tr>
<tr>
<td>QRSN (_N_2)</td>
<td>Query Restricted Segment Net (global)</td>
<td>0.41</td>
<td>2.71 \times 10^{-5}</td>
<td>7.22</td>
<td>4.46</td>
<td>2000</td>
<td>-2.01</td>
<td>-3.25</td>
</tr>
<tr>
<td>NUW (_N_1) ( [2] )</td>
<td>NL Unrestricted Word Net (local)</td>
<td>0.69</td>
<td>1.55 \times 10^{-4}</td>
<td>2.63</td>
<td>3.03</td>
<td>\approx 2500</td>
<td>-1.50</td>
<td>-2.70</td>
</tr>
<tr>
<td>NRWN (_N_1) ( [2] )</td>
<td>NL Restricted Word Network (local)</td>
<td>0.44</td>
<td>1.55 \times 10^{-4}</td>
<td>2.67</td>
<td>3.06</td>
<td>\approx 2500</td>
<td>-1.50</td>
<td>-2.70</td>
</tr>
</tbody>
</table>

---

### Figure 1: Cumulative degree distributions for segment networks (QUUSN\(_N_2\) and QRSN\(_N_2\)).
and Web search queries. While single regime power-law networks are abundant in nature and their emergence can be explained by the law of “preferential attachment” (see [1] for definition), two-regime power-law networks are quite rare. In comparison, query networks are not only similar to word co-occurrence networks in terms of the qualitative behavior of their DD’s, even the values of the exponents are very similar. Thus, we are tempted to conclude that queries, at least in some respects, are very similar to natural languages. In the next section, we discuss the common and different traits between natural language and query networks, explaining their linguistic and socio-cultural significance.

2. INSIGHTS FROM THE NETWORKS

In [2], the topological properties of word co-occurrence networks have been explained in terms of well-known linguistic theories. In absence of any prior linguistic study of Web search queries, here we make an attempt to explain the similarities and differences between query and natural language networks by drawing analogy to linguistic theories and their applicability in the context of queries.

2.1 Deviation from Small-world Property

Small-world networks are found in many natural, technological and social systems and have been shown to exhibit interesting properties [5]. Cancho and Solé show that word co-occurrence networks are small-world and proposes that their linguistic and socio-cultural significance.

2.2 Kernel and Peripheral Lexicons

The emergence of two regimes in the DD of word co-occurrence networks has been explained by the existence of distinct kernel and peripheral lexicons. Extending this analogy, one can argue that the language of Web search queries is characterized by the existence of two different types of words or segments (henceforth referred to as units): Units popular in queries acquire high degrees and constitute the kernel lexicon (K-Lex), while rarer ones, with degree much lower than those in the kernel, form the peripheral lexicon (P-Lex). $k_{max}$ is observed to be the same for unrestricted and restricted network pairs, which is also observed for Standard English [2].
properties of co-occurrence networks for natural language (Standard English) and those for queries. We highlight these individually in the subsequent paragraphs.

The kernel-to-periphery ratio is much higher in queries than that for English. For Standard English, the ratio of the sizes of P-Lex to K-Lex is $\approx 85$, while for queries it ranges between 250 and 850 (average kernel and periphery sizes are 1285 and 648615 respectively). The presence of a very large periphery means that in a query, one is highly likely to encounter previously unseen words, something which is much less probable in natural language. This is because users continuously try to provide “discriminating” (instead of popular) words and phrases to narrow down the search space.

Periphery stands by itself. A large diameter ($d \approx 7$) for query networks implies that one must also traverse intra-peripheral edges to reach one peripheral unit from another. Interestingly, such linkages emerge from fail queries. These units may not have a direct connection to the kernel and hence do not provide the search engine with enough known context when first encountered. This makes processing of rare queries harder than processing of new words in natural language sentences. This is because, new words always occur in the context of known words that provide enough context for word understanding. For example, in query networks, we have the following shortest path: airedale terrier $\leftrightarrow$ tumor $\leftrightarrow$ where $\leftrightarrow$ download $\leftrightarrow$ prison break. Here the first two and the last segments are from the periphery and the rest from the kernel. Clearly, the query constituted of the first two segments “airedale terrier” “tumor” is hard to understand and process automatically.

The kernel network is much less tightly coupled than in English. The kernel network (KN) is defined to be formed by nodes from the kernel lexicon only, and the edges that go between them [2]. The CDD of the KN for query networks falls off as power law (Fig. 2) with exponent $\gamma_{KN}$ ranging between $-2.74$ and $-2.96$, and the average clustering coefficient is 0.59 (fully connected networks will have a CC of 1); thus, unlike word-occurrence networks for natural language, KN for queries does not exhibit a clique-like behavior. In natural language, sentences are long, and contain stopwords like the and in the same sentence. This is primarily responsible for the creation of cliques in the kernel network. However, queries are mostly devoid of stopwords, and modifiers belonging to different classes rarely occur in the same query — for instance, top modifiers like pdf and imdb do not share an edge.

The peripheral network consists of a large number of small disconnected components. The degree distribution in the peripheral network (PN, defined analogously as KN) decays much more slowly (inset, Fig. 2), due to the presence of a large number of approximately equal-sized disconnected components. The largest component of the PN contains $\approx 20000$ nodes, which, however, is very large compared to the corresponding network in natural language.

Most edges run between kernel and periphery for queries, whereas intra-kernel edges dominate in natural language. From the distribution of the types of edges ($\approx 98.76\%$ between kernel and periphery, $\approx 0.48\%$ intra-kernel, $\approx 0.76\%$ intra-periphery), we can see that the high degrees in the kernel are significantly due to contributions from peripheral units. As mentioned earlier, a typical query generally has a rare unit as a distinguishing feature.

Socio-cultural factors govern the Kernel-Periphery distinction in queries. While for natural language the divide between the K-Lex and P-Lex emerges primarily due to cognitive/syntactic and only marginally due to socio-cultural factors, for queries, the division emerges mainly due to socio-cultural factors and only marginally due to syntactic factors. For instance, *lyrics* is in the K-Lex and *genes* in the P-Lex only because lyrics search is far more popular than gene search; this is purely a socio-cultural phenomenon and has nothing to do with the syntactic structure of the query language. However, the fact that Australian cities, *download* or *wiki* are in K-Lex is more an effect of the structure of queries; though strictly speaking, almost everything about queries is socio-cultural; there is hardly anything cognitive here. Query logs are a representation of the collective information need of the people of a particular geography and demographics.

3. CONCLUSIONS AND FUTURE WORK

The contribution of this paper lies in establishing that word co-occurrence patterns in Web search queries and natural languages are strikingly similar, yet very different from each other. While like natural language, queries reflect a Kernel-Periphery distinction implying an underlying linguistic structure, at the same time, unlike natural languages, query networks lack small-world property which has been argued to be essential in natural language for quickly retrieving words from the mind. More importantly, the underlying implications of these differences are that it is much more difficult to understand the context of a word or a segment in a query. A tightly-knit kernel connected to every peripheral node forms the basis for contextual inferencing in natural language both by man and machine. In queries, the absence of such a kernel and a much larger periphery results in a high surprise factor for several peripheral units. The capability of peripheral units to exist by themselves makes POS identification hard in queries. Understanding these fundamental differences may open up an area where researchers would more formally characterize query grammar. We believe that the application of complex networks to query representation holds immense potential, of which this work is just a beginning.

4. REFERENCES


