The Semantic Structure of Query Search Results

Ying Liu

1Department of Computer Science, Mathematics and Science, College of Professional Studies, St. John’s University, Queens, NY 11439
liuy1@stjohns.edu

Abstract - Organization of query’s search results is a key issue of modern information retrieval systems. In recent years, ranking and listing Web pages based on their Web links have made great success in Web information systems. However, just ranking documents retrieved is not sufficient to users and this kind of ranking is only specific to text collections with extra linkage information among documents. Moreover, the organization of search results relies not only the extra linkage information, but also the intra content information. In the case of bad-quality extra linkage information, the understanding of intra content information of documents retrieved plays the crucial role. Therefore, to effectively organize query’s search results, in this paper, we studied the semantic structure, i.e., the content-based relationships, of query’s search results from the perspective of statistical networks. We showed that such semantic structure is a complex network with the properties of small worlds, scale free and hierarchy. Therefore, the semantic relationships between documents retrieved are revealed and hence provide a beneficial understanding of how these documents can be organized for users according to their contents.

Keywords: Query Search Results; Semantic Network; Small Worlds Network; Scale Free Network

1 Introduction

In the last decade, the online text in the Internet has grown into a massive repository of information. Therefore, information retrieval increasingly attracted more interests and attention with the rapid growth of the Internet. The vast volume of online text information has made the phenomenon “information overload” worse in modern information retrieval systems. For instance, an ad-hoc query could return too many results, only few of which are relevant to users. Therefore, the bulk of studies in modern information retrieval have focused on how to present relevant documents effectively, e.g., ranking Web pages by exploring their hyperlinks, or grouping documents into clusters with distinct topics. With hyperlinks linking Web pages, Web information retrieval regards the Word Wide Web as a graph and explores the link structure of the Web graph to help measure the relevance of each retrieved Web page. The Web graph has been reported to be a complex network with remarkable properties [1, 2], small-world, scale-free, hierarchy, community structure and so on. These discoveries have potentially advanced the development of modern information retrieval techniques [3, 4].

However, in recent years, with intense and considerable interests in exploring and discovering the link structure of the Web, there are few researches paying attention to the study of the semantic structure of query search results. Recalling popular ranking algorithms in Web information retrieval, e.g., HITS [5] and PageRank [6], they firstly retrieve large amounts of query search results containing query’s text information, and then explore their link structure, which is a network with documents as nodes and their hyperlinks as edges. Similarly, the semantic structure of query search results is a network of content relationships among search results, with documents as nodes and their textual relationships as edges. Like the exploration of the link structure of query search results, understanding and analyzing the semantic structure of query search results can also effectively enhance the organization and representation of search results. More important, unlike the Web, many other text resources are free text and do not have the additional linkage information to be exploited, e.g., digital libraries. In these cases, the understanding of their semantic structures is important to information retrieval systems. Although there have been much researches in the topological properties of the link structure of the Web, there are few studies related to the semantic structure. For example, Menczer [7] studied the topological relationship of the link structure and the semantic structure in the Web. The power-law relationship was discovered, and then based on this discovery, two applications were discussed: a content-based generative model for explaining Web growth and content-based crawling algorithms for Web navigation. However, Menczer’s work did not involve with the study of the semantic structure and his results are limited to the Web. In contrast, the study of purely textual relationships among query search results has potential impact on all kinds of text resources, including free-text collections without linkage information.

Therefore, in this paper, given a query, we present a study of a semantic network of this query’s search results that is large – possibly containing tens of thousands of retrieved documents – and for which a precise definition of semantic proximity is possible. This network is weighted, where the weight of a semantic link between two documents is their content similarity.
2 Semantic Networks of a Query Search Results

We study networks of a query search results (or retrieved documents) in which two retrieved documents are considered connected if their similarity in text content is high enough to exceed a value. There are two factors in this definition affecting the construction of semantic edges among search results. One is the definition of content similarity purely based on textual information of retrieved documents. In this paper, we use the cosine similarity function, which is a widely used measure in information retrieval, as the content similarity measure. It is defined as follows,

\[ s(d_1, d_2) = \frac{\sum_{j \in d_1 \cap d_2} W_{d_1j} W_{d_2j}}{\sqrt{(\sum_{j \in d_1} W_{d_1j}^2)(\sum_{j \in d_2} W_{d_2j}^2)}} \]

where \( W_{id} \) is some weight function for term \( j \) in the retrieved document \( d \), e.g., term frequency (TF) or term frequency-inverse document frequency (TFIDF) function. Hence, similarity values are in the range of zero (indicating completely not similar) and one (meaning the same). The other is the similarity threshold defining if two retrieved documents are similar enough to be connected. To investigate the impact of similarity thresholds on the topology of semantic networks, we use different similarity thresholds by iterating the range of \([0,1]\) with a small gap.

In addition to the introduction of content similarity function and similarity thresholds, the preprocessing of text documents follows typical steps in information retrieval: each retrieved documents are broken into tokens, which, in this paper, single terms (or words). Word stemming is used to truncate suffixes so that words having the same root (e.g., activate, activates, and activating) are collapsed to the same term for frequency counting. In our work, Porter’s stemmer was applied [8]. Stop lists are typically used to filter out non-scientific English words. The standard TFIDF function was used, in this paper, to assign the weight to each term in the document. Then each document was modeled as an \( m \)-dimensional TFIDF vector, where \( m \) is the number of distinct terms in all the search results. Formally, a document is a vector \((t_{idf_1}, t_{idf_2}, \ldots, t_{idf_m})\), where \( t_{idf_i} \) is the \( t_{idf} \) value of word \( i \). Then a term-by-document matrix was built, in which, each column represents a document, and each row represents a term. The values in the matrix are TFIDF weights. If a term does not appear in one document, then a value 0 is assigned to that cell in the matrix. To summarize the process of constructing semantic networks of query search results, a flowchart is presented in Figure 1.

In this paper, we analyze the semantic networks constructed from queries’ search results returned by three different online information retrieval systems, i.e., a large digital library - PubMed\(^1\), a small digital library Human-Computer Interaction (HCI) Bibliography\(^2\) (both of these two libraries provide abstracts of retrieved research papers as textual information), and a popular Web search engine - Google\(^3\) (Web pages are retrieved by removing HTML tags and therefore textual information remains after preprocessing). To diversify queries, we consider, specific queries, i.e., a set of ten cancers’ names in PubMed corresponding to ten categories from Society of Surgical Oncology (http://www.surgonc.org) Annotated Bibliography (Among them, we selected only the query “breast cancer” which returns the largest number of relevant documents, while results of other nine cancer queries are reported in the supplementary information of this paper.); and a broad-topic query, i.e., “cognitive” in HCI Bibliography (this term is commonly used in HCI). Moreover, to we also submit a cancer name “breast cancer” as a specific query to the Google and analyze their semantic networks of the first 1000 ranked Web pages (after excluding PDF format files’ links in the search results and the secured web pages, the number of actual Web pages used in the experiment are less than 1000.). The summary of data used is listed in Table 1.

Table S1. Summary of queries’ search results

---

\(^1\) PubMed is a service of the U.S. National Library of Medicine that includes over 16 million citations from MEDLINE and other life science journals for biomedical articles back to the 1950s. Its address is http://www.ncbi.nlm.nih.gov/entrez.

\(^2\) HCI Bibliography is an online resource for academic use in Human-Computer Interaction, which provides searching service over 36,000 publications about Human-Computer Interaction. Its address is http://www.hcibib.org.

\(^3\) Google is one of the most popular Web search engine in the World Wide Web. Its address is http://www.google.com.
3 Results

Following the flowchart of constructing semantic networks in query’s search results in Figure 1, we set similarity thresholds as 0.8, 0.6, 0.4, 0.2 and 0.1, and then five semantic networks are obtained for each query’s search results. In the following, we investigated structural properties of these semantic networks.

3.1 Clustering Coefficient

The clustering coefficient was first introduced by Watts and Strogatz [9] as a graph metric to determine if a network is highly clustered or not. Let a network with \( n \) nodes be \( G(V,E) \) with \( |V|=n \). On a selected node \( i \), it has \( k_i \) edges which connect it to other nodes. If the first neighbors of the original node are part of a clique, there would be \( \frac{k_i(k_i-1)}{2} \) edges between them. Therefore, the ratio between the number \( e_i \) of edges that actually exist between these \( k_i \) first neighbors and total number \( \frac{k_i(k_i-1)}{2} \), yield the value of clustering coefficient of node \( i \),

\[
C(i) = \frac{2e_i}{k_i(k_i-1)}
\]

Consequently, the average of all nodes’ clustering coefficient in \( G \) is the overall clustering coefficient of \( G \), i.e.,

\[
C(G) = \frac{1}{n} \sum C(i).
\]

We called this computation method of clustering coefficient as local clustering coefficient (LCC), because it is an average of each node’s clustering coefficient locally computed. Similarly, an alternative definition of clustering coefficient was proposed by Newman [10]. It is globally computed through counting the number of triangles and 2-length paths and therefore called global clustering coefficient (GCC). It is formally defined as follows,

\[
GCC(G) = \frac{6 \times \text{number of trianlge in the network}}{\text{number of path of length two}}
\]

Table 2. Summary of Query “Breast Cancer” Searched in PubMed (47871 nodes/documents)

<table>
<thead>
<tr>
<th>Searched in PubMed</th>
<th>Query</th>
<th>#Documents Retrieved</th>
<th>#Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer</td>
<td>47871</td>
<td>4087</td>
<td></td>
</tr>
<tr>
<td>Colorectal Cancer</td>
<td>26556</td>
<td>2940</td>
<td></td>
</tr>
<tr>
<td>Searched in HCI</td>
<td>Cognitive</td>
<td>2310</td>
<td>6359</td>
</tr>
<tr>
<td>Searched in Google</td>
<td>Breast Cancer</td>
<td>788</td>
<td>1733</td>
</tr>
</tbody>
</table>

3.2 Average Shortest Path Length

The average shortest path length, which is also called “characteristic path length” by Watts and Strogatz [9], measures the typical separation between two nodes in a network. It is defined as

\[
L(G) = \frac{1}{|P|} \sum_{i \in P} d_{ij}
\]

where, \( P \) is the set of all possible shortest paths in the network \( G \), and \( |P| \) is thus the number of shortest paths in \( G \).

3.3 Small-World Effect

The introduction of the small-world network by Watts and Strogatz [9] implicated that most real-world networks lie between regular networks and random networks, by sharing the same phenomenon: they are highly clustered, yet have small separation of nodes. More often than not, the clustering degree is measured by LCC or GCC, and the separation of nodes is measured by the average shortest path. Typically, identifying small world networks follows the framework of Watts and Strogatz by comparing real-world networks \( G_{\text{actual}} \) with random networks \( G_{\text{random}} \) with the same number of nodes and edges in terms of clustering coefficient and characteristic path length, i.e., if the actual graph \( G_{\text{actual}} \) satisfies \( C(G_{\text{actual}}) >> C(G_{\text{random}}) \) and \( L(G_{\text{actual}}) \approx L(G_{\text{random}}) \), it is a small world network. In Table 2, 3, 4 and 5, values of these measures in semantic networks of four queries’ search results are summarized.

Observing semantic networks of these four queries’ search results, we found an interesting phenomenon: “when the similarity threshold is high (e.g. \( \sigma =0.8 \)), there are small number of edges (even much smaller than the number of nodes) in the semantic network and therefore do not appear as small-world networks, although most of these semantic networks have larger clustering coefficients (e.g., in Table 4, \( LCC(G_{\text{actual}})=0.912 \), which is much larger than \( LCC(G_{\text{random}})=0 \)) than random networks and approximately smaller average shortest path length (their ratio is less than 5 times); however, when the similarity threshold is set lower (e.g., 0.2 and 0.1), the semantic networks do show the small-world phenomenon with \( L(G_{\text{actual}}) \approx L(G_{\text{random}}) \) but \( C(G_{\text{actual}}) >> C(G_{\text{random}}) \).”
4 Conclusion

In this paper, we investigated the semantic structure or content relationship of search results of queries which can be domain-specific or broad-topic. The results showed that the semantic structures of query’s search results do show properties of complex networks like most real-world networks. In specific, the semantic networks of query’s search results are small-world, scale-free, and hierarchy. We believe these results can provide benefits and useful implications to the design of information retrieval systems for better organization and performance of search results.

5 References