On Topology of Baidu's Association Graph Based on General Recommendation Engine and Users' Behavior

Cong Men, Wanwan Tang, Po Zhang and Junqi Hou
Baidu Inc.
mencong@baidu.com

ABSTRACT

To better meet users' underlying navigational requirement, search engines like Baidu has developed general recommendation engine and provided related entities on the right side of the search engine results page(SERP). However, users' behavior have not been well investigated after the association of individual queries in search engine. To better understand users' navigational activities, we propose a new method to map users' behavior to an association graph and make graph analysis. Interesting properties like clustering and assortativity are found in this association graph. This study provides a new perspective on research of semantic network and users' navigational behavior on SERP.

Categories and Subject Descriptors

E.1 [**DATA STRUCTURES**]: Graphs and networks; H.2.8 [**Database Applications**]: Data mining; I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic networks

Keywords

Graph analysis, General recommendation engine, User behavior analysis

1. INTRODUCTION

In order to better satisfy users' needs and further stimulate users' underlying requirement, search engines(Google, Bing, Baidu etc.) tend to recommend related entities on the right side of search engine results page(SERP) based on the technique of knowledge graph in recent years. For example, when the query is "running man"(it is a popular tv show in China, here we directly translate the query into English.), general recommendation engine will display the most related actors and tv programs on the right-side of SERP as shown in Fig.1.

Isolated queries become associated with each other, and could be seen as a huge graph due to this technique. User experience may closely relate to the topology optimization of the network. Research on topology of the huge graph will provide new perspectives on the optimization of knowledge graph.

Figure 1: An example of Baidu right-side recommendation.

2. DATA AND METHODOLOGY

We divide searches into two types based on their origins. One type is defined as information-oriented, which means user type a query in the search box, and obtain the corresponding information. The other is defined as navigation-oriented, which means user click the recommended entity on the right side, and tend to obtain additional information of the new entity. We will compare users' behavior of these two types after mapping them into graphs.

We construct the directed graphs based on users' click-through data of Baidu right-side recommendation on SERP for these two types of searches. We define these two types of graph as informational graph (IG) and navigational graph (NG). Vertices in the graphs are queries, which are different from the query-url bipartite graphs in Ref.[1]. If the click-through rate for a $< q, e >$ pair is above a threshold, an edge from vertex q to vertex e exists. Here, q represents the query in the search box, and e represents the displayed entity on the right side of SERP. It makes the relevance of query q and e more reliable by selecting edges based on users' behavior. Here, we select queries whose average search volume per day are more than 50 to construct the graph. Ten days' click-data of all users' search behavior in Baidu search engine are chosen as the original dataset.

Then, we apply network analysis metrics like clustering coefficient, degree assortativity to investigate the topologies of the informational and navigational graph. We aim to

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742724.

Table 1: Basic graph metrics of Baidu association graph

Metrics	IG	NG
Size	471738	54732
Connectivity	0.101	0.863
Clustering Coefficient	0.162	0.302

Table 2: Assortativity coefficient of Baidu association graph

Metrics	IG	NG
Source:in,Destination:in	0.0188	0.092
Source:in,Destination:out	0.075	0.139
Source:out,Destination:in	0.116	0.149
Source:out,Destination:out	0.401	0.479

unveil the underlying characteristics of informational and navigational search patterns by mapping the users' behavior to a large graph.

3. GRAPH ANALYSIS

Several basic graph metrics are displayed in Table.1. The size of navigational graph is much smaller than that of informational graph. Because most users search for specific information in search engine not for browsing. Furthermore, the objective of Baidu's right recommendation is to meet users' extended demand instead of the master demand.

The largest strongly connected component of navigational graph contains 86.3% of all queries, which means users can move from one query to another only by clicking right recommended entities in this subset of 86.3% queries in the graph. It gives users a convenient way to browse for related information. Here, this connectivity number is lower than the traditional social network like Facebook and larger than Twitter[2, 3]. Because there exist some isolated search demands in the navigational graph. In informational graph, the largest strongly connected component contain 10% queries. Moreover, users tend to focus on the left-side of SERP and neglect the recommended entities in informational search.

To better quantify how tightly queries are connected, we apply shortest path length to represent the minimum numbers of right clicks users need from one query to another. The average shortest path length of the largest connected component in navigational graph is 8.6, which is higher than that of traditional semantic network like thesaurus graph or word associative graph[4]. It is also higher than that in traditional social network[2, 3]. One reason is that the connections in our graph is usually not reciprocated. Another reason is that only the most related entities are recommend on the right-side of SERP, which makes the out-degree of each node has a upper bound. That is one of the main differences between our network and other complex networks.

The clustering coefficient is used to quantify the fraction of queries whose correlated queries are themselves correlated in this graph[5]. Clustering coefficient in navigational graph is 0.302, which is larger than that of informational graph 0.162. These metrics are lower than that of thesaurus graph but close to that of word associative graph[4].

Degree assortativity measures the similarity of connections in the graph with respect to the node degree. Here, both in-degree and out-degree correlations of the graph are calculated. We consider four types of degrees: source in-degree (SID), source out-degree (SOD), destination in-degree (DID), and destination out-degree (DOD).

In-degree and out-degree correlations of these two types of graph are shown in Table.2. Here, in-degree of node depicts the number of pre-queries of this query in this graph, and it measures the popularity of the query. Out-degree of node quantify to what extent the query could trigger users' new navigational demands. This metric has a upper bound because space for right recommendation is limited. It is found that all of these metrics are positive. Assortativity coefficient of SOD and DOD is shown to be the highest. That means navigation-oriented queries tend to trigger more navigational searches. Furthermore, Popular queries may trigger users to search more related queries by right-side clicks in navigational graph. However, this phenomenon is not obvious in informational graph. Furthermore, popular queries' recommended queries tend to be popular and trigger more related navigational searches. Navigation-oriented queries also tend to recommend popular queries to users. It is also interesting that all assortativity coefficients in navigational graph are higher than those in informational graph.

4. DISCUSSION AND CONCLUSIONS

We propose a new semantic graph based on Baidu knowledge graph and users' behavior in this research. Interesting properties like clustering and assortativity are notably observed in these association graphs especially in navigational graph. This analysis could provide more insights on users behavior and navigational patterns beyond the traditional metrics like click-through rate. It may also provide new characteristics to improve the quality of search engines' right recommendation. Furthermore, different properties in these graphs are also found after comparing with those in traditional social networks and semantic networks.

As a new semantic graph, this search association graph has some advantages than traditional semantic networks. (1) It is convenient to built a even larger graph after processing more days of search log. (2) Some underlying association rules are easier to be found based on users' behavior rather than based on their literal meanings. Therefore, analysis of users' behavior in search engine provide a new reliable method to build large semantic network.

5. REFERENCES

[1] R. Baeza-Yates, A. Tiberi. Extracting Semantic Relations from Query Logs. Conference on Knowledge Discovery and Data Mining, 2007.

[2] S. A Myers, A. Sharma, P. Gupta and J. Lin. Information graph or social graph?: the structure of the twitter follow graph. In International Conference on the World Wide Web, 2014.

[3] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. Nature, 1998.

[4] J. Borge-Holthoefer, A. Arenas, Semantic Networks: Structure and Dynamics. Entropy 2010.

[5] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the Facebook social graph. arXiv, 2011.