The Web without search engines

Alexander Ponomarenko

National Research University Higher School of Economics, Laboratory of Algorithms and Technologies for Network Analysis, 136 Rodionova str., Nizhny Novgorod 603093, Russia

Abstract

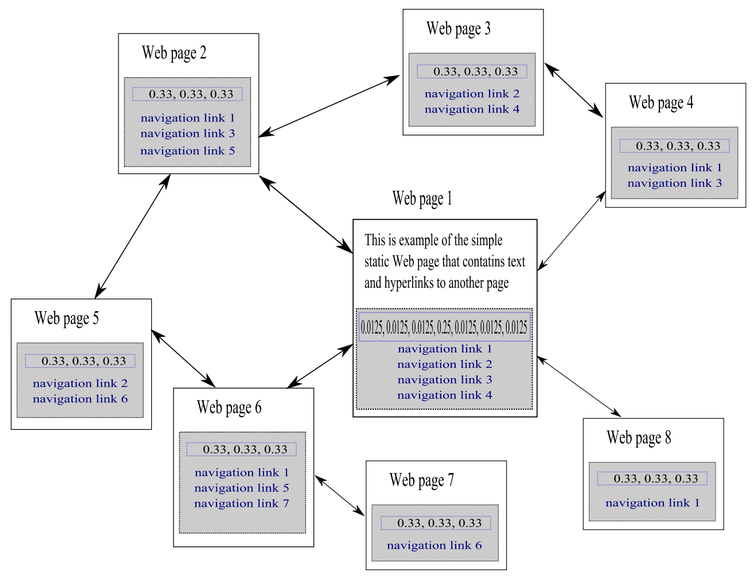
A standard approach for searching data distributed across the Internet is to use a search engines which build index for this purpose. Information can have any form such as documents, HTML pages or RDF triples. An index can be stored in a cluster or can be deployed to distributed environment using DHT. In any case the two copies of data have appeared.

It results in two problems. The first problem is that the volume of all information in the Internet can be significantly larger than the index capacity of a search engine. The second problem is that information in the index becomes irrelevant over time and the index should be rebuilt.

In this work we propose a novel approach how to organize information for further search without duplicating it and without building a traditional index.

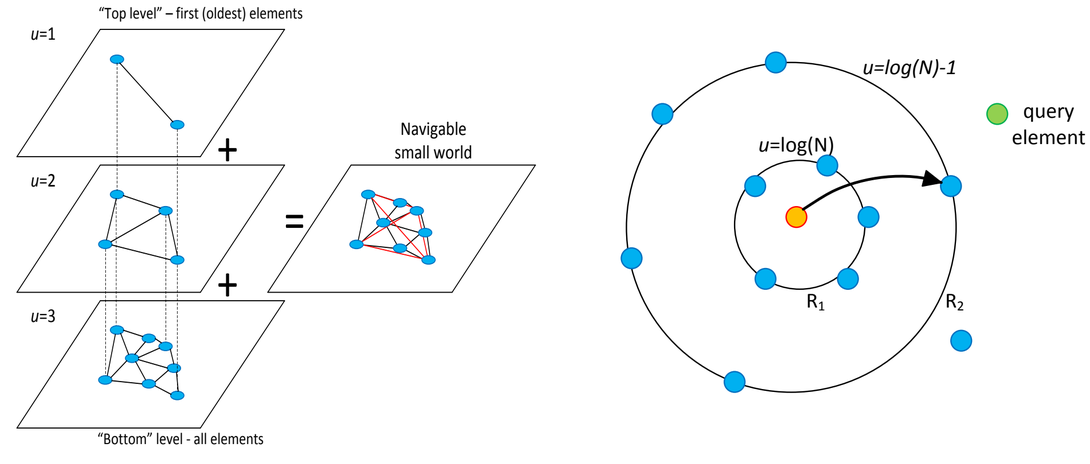
We suggest that information should be organized into a data structure by building an overlay network on the level of the separate information objects such as HTML pages or RDF triples, based on the similarity between them.

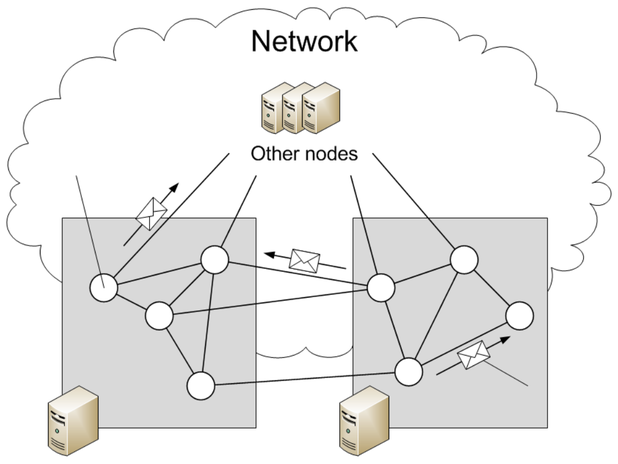
For this purpose the recently developed data structure with small world properties designed for search in high dimensional metric spaces was applied.



Core Idea

The core idea is that search of the most relevant page in the Web to the query *q* can be considered as nearest neighbor search in metric space *M(U, d)*, where *d* is a distance function which corresponds to relevance function and *U* is a set of all possible Web pages. The every Web-page can be considered as an independent node of peer-to-peer system with it's own routing table. The network graph of this peer-to-peer system can be build so that the efficient nearest neighbor search can be done greedy forwarding query messages from one node to node. We propose to build the network graph by using construction algorithm of Metrized Small World data structure [1], [2] over the finite set $X \subset U$





The Metrized Small World Data Structure

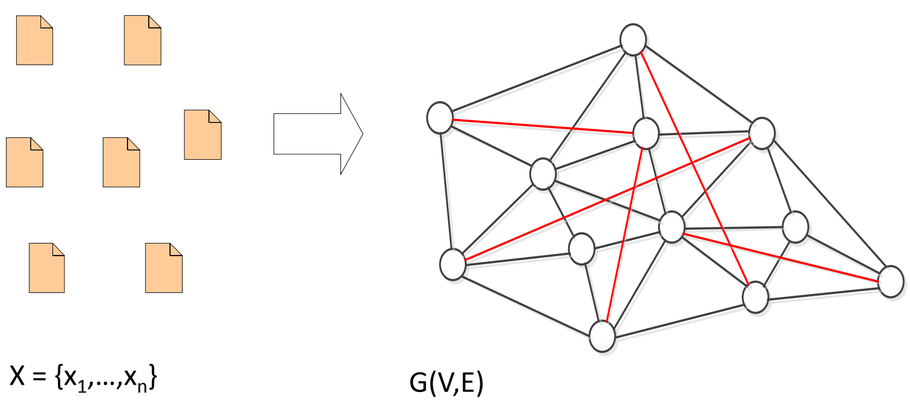
The data structure S is constructed as a navigable small world network represented by a graph G(V,E), where objects from the set X are uniquely mapped to vertices from the set V. The set of edges E is determined by the structure construction algorithm. Since each vertex is uniquely mapped to an element from the set X, we will use the terms "vertex", "element" and "object" interchangeably. We will use the term "friends" for vertices that share an edge. The list of vertices that share a common edge with the vertex v_i is called the friend list of the vertex v_i.

The data structure use a variation of the greedy search algorithm as a base algorithm for the k-NN search. It traverses the graph from an element to another element each time selecting an unvisited friend closest to the query until it reaches a stop condition.

It is important to note that links (edges) in the graph serve two distinct purposes:

1) There is a subset of short-range links, which are used as an approximation of the Delaunay graph required by the greedy search algorithm.

2) Another subset is the long-range links, which are used for logarithmic scaling of the greedy search. Long-range links are responsible for the navigation small world properties of the constructed graph.



Computational Experiments

In order to verify of assumptions we have performed the computational experiment. As Web-pages we have used Wikipedia dataset. Wikipedia is dataset that contains 3.2 million vectors represented in a sparse format. Each vector corresponds to the frequency term vector of the Wikipedia page. This set has an extremely high dimensionality (more than 100 thousand elements).

$$d_j = (w_{1,j}, w_{2,j}, ..., w_{n,j})$$
$$q = (w_{1,q}, w_{2,q}, ..., w_{n,q})$$

$$sim(d_j, q) = \frac{d_j \cdot q}{\| d_j \| \cdot \| q \|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \sqrt{\sum_{i=1}^{n} w_{i,q}^2}}$$

References

[1] Malkov, Y., Ponomarenko, A., Logvinov, A., & Krylov, V. (2013). Approximate nearest neighbor algorithm based on navigable small world graphs. *Information Systems*.

[2] Пономаренко А. А., Мальков Ю. А., Логвинов А. А., Крылов В. В. Структура со свойствами тесного мира для решения задачи поиска ближайшего соседа в метрическом пространстве // Вестник Нижегородского университета им. Н.И. Лобачевского. 2012. № 5. С. 409-415.

ИТиС 2014

NATIONAL RESEARCH UNIVERSITY

Email: aponomarenko@hse.ru