Wowd distributed search engine

Computers in Scientific Discovery 5

Aleksandar Ilić
aleksandari@gmail.com
University of Niš, Serbia

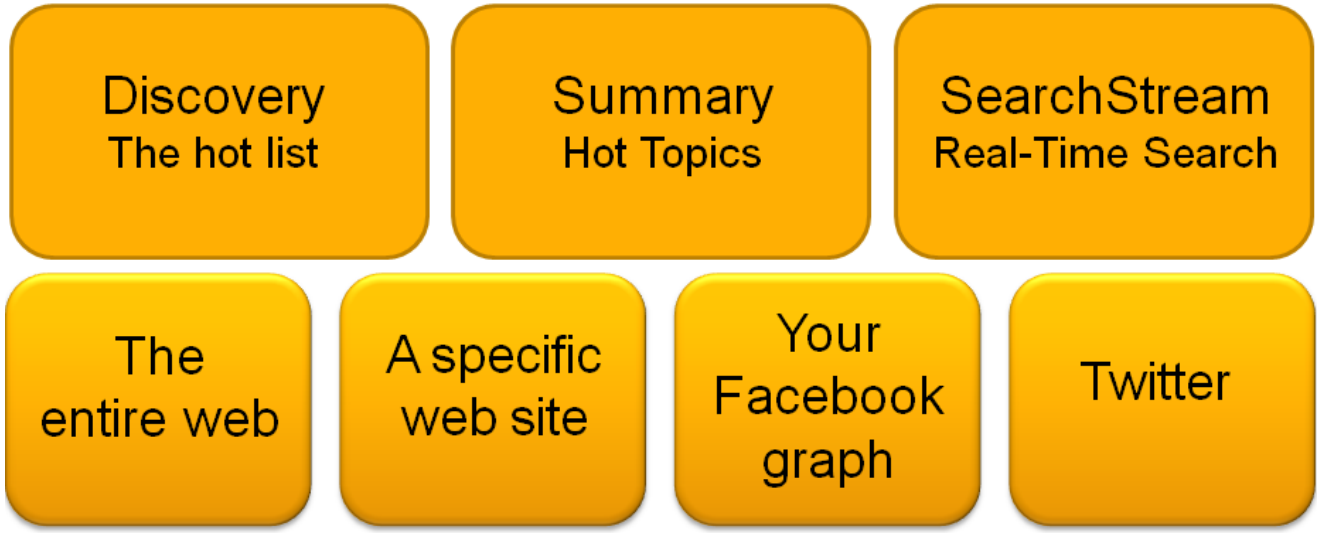
Sheffield, July 2010

- Wowd
  - Distributed P2P real-time discovery & search engine
  - http://www.wowd.com/
- Graphs in Wowd
  - routable graphs
  - ranking in internet graph
  - ranking in social graph

Background

- Founded by Borislav Agapiev in 2007
- Development team is completely in Serbia (JAVA)
- Investors are USA venture capital firms
  - Draper Fisher Jurvetson, KPG Ventures, Stanford University
- Research in many cutting-edge fields
- **Studying topology and traffic of large-scale networks**

What is Wowd?


Discovery
The hot list
Summary
Hot Topics
SearchStream
Real-Time Search
The entire web
A specific web site
Your Facebook graph
Twitter

Age of Information

Finding meaning in unstructured data requires using different techniques:

- **Google's PageRank** - finding the relative importance of web pages for searching.
- **Social Network Analysis** - finding how groups are divided, who is the most popular and who hangs out with who…
- **Bioinformatics** - find which proteins function similarly.
- **Pattern Matching** - given a pattern find all the instances of a subgraph of this pattern.

Reference search vs. Real-time discovery

Google: reference search

I am looking for information on X

(1) Think of something
(2) Go to Google, type it in, hit enter
(3) Look through the results, refine query as needed

Wowd: discovery in real-time

I am watching for developments (in X)

(1) Wonder what's going on
(2) Go to Wowd, look at the Hot List, Hot Topics
(3) Click on a topic of interest, watch new material roll in

Graphs in Wowd

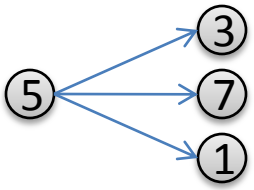
- construction of routable graph of computers
  - millions of vertices
- ranking in internet graph
  - from 100 million to tens of billion of vertices
- ranking in social graph
  - 10-100 million of vertices
- graphs in bioinformatics
  - from 100 vertices to 100 million of vertices (proteins, molecules, atoms)

Routable graphs

- set of nodes (computers) in a distributed network
- how can any node get to any other node
  - as fast as possible
- create an algorithm for constructing a graph

Routable graphs

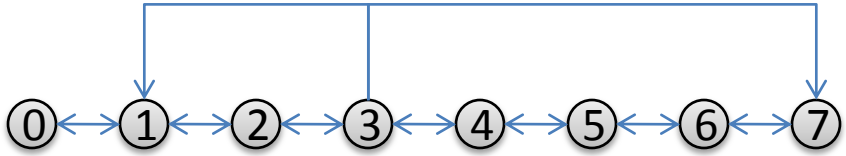
- vertices are labeled
  - random binary 64bit number
- directed
- routable
  - must be possible to find a path to any label
  - labels of neighbors (only) are known



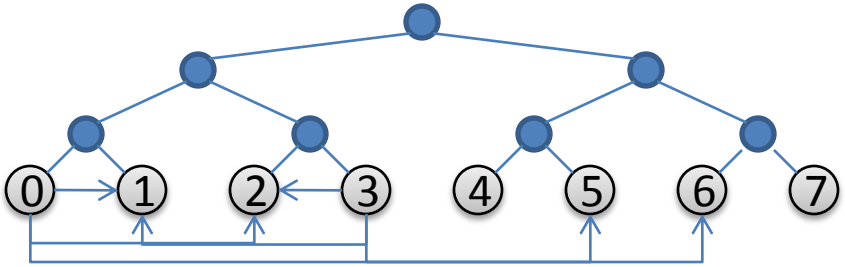
path from 5 to 4?

Routable graphs

- structure must be defined
  - ordering:
    - each vertex must have connection to first lower and first higher
    - skip lists:



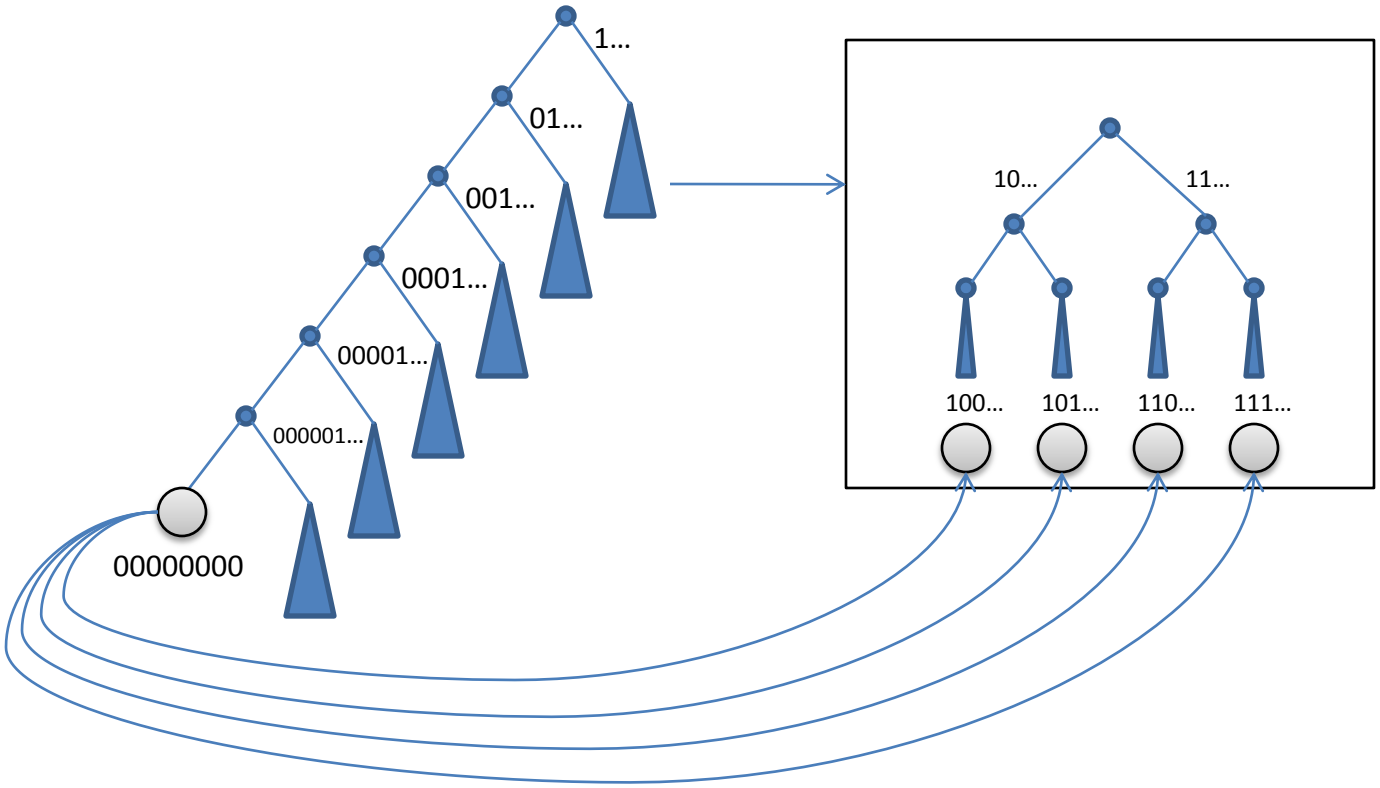
  - distance:
    - for any label, each must have connection to at least one with closer label
    - XOR distance:



Routable graphs

- routable k-connected
  - only findable paths are considered
- Dynamic
  - adding and removing vertices, while keeping requirements
  - locality of change
  - adding vertex (only edges to and from it can be added)
  - removing vertex (only edges instead of removed ones are allowed)
- degree of nodes is limited
  - maintenance limit

Routable graphs



Routable graphs – in numbers

|V(G)|	Max degree	Average distance	Theoretical optimum	Average/Theor.
2^{10} (1K)	191	1.89	1.81	1.04
2^{15} (32K)	351	2.77	1.99	1.39
2^{20}(1M)	511	3.62	2.75	1.32
2^{22} (4M)	575	3.93	2.92	1.35
2^{24} (16M)	639	4.29	2.98	1.44

Note: theoretical optimum with respect to only max degree constraint

Degree/diameter problem

- Given natural numbers Δ and *D*, find the largest possible number of nodes $n_{\Delta,D}$ in a graph of maximum degree Δ and diameter *D*.
- Moore bound:

$$n_{\Delta,D} \leq 1 + \Delta + \Delta(\Delta-1) + \Delta(\Delta-1)^2 + \ldots + \Delta(\Delta-1)^{D-1}$$

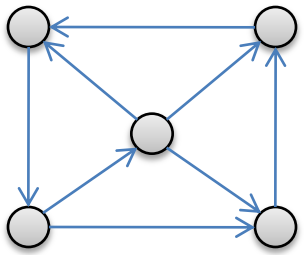
- **Open question**: Does there exist a Moore graph of diameter 2 and degree 57?

Ranking in internet graph

- set of internet pages
- structure – links between them
- how to rank/sort them?

Ranking in internet graph

- random surfer model



- rank of pages = probability on being on each page
- if A is adjacency matrix, it becomes:

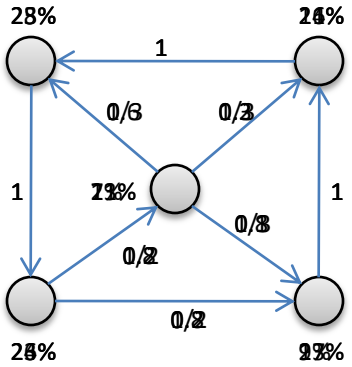
$$r = \lambda A r + (1 - \lambda)$$

- converges if sum of each row is ≤1
- solution is largest eigenvalue

Ranking in internet graph

Edge weights:

- uniform $e(u,v) = \frac{1}{|N(u)|}$
  - Google's PageRank



- actual probability of surfer following that link
  - ours EdgeRank (patented)
  - simplified: count clicks on each link, and use:

$$e(u,v) = \frac{c(u,v)}{\sum_{t \in N(u)} c(u,t)}$$

Ranking in internet graph

Distributed iterative calculation

- number of needed iterations is small
  - initial: 5-10 iterations
  - new pages: 2-3 iterations
- $O(n_{iter} E(G))$ and trivially distributed

Ranking in social graph

- set of social users
  - Twitter users
    - graph publicly available
    - directed social graph
- how to rank/sort them?
  - needed to best use attention frontier
- same idea – random walk

Applications

- **Global alignment of multiple protein-protein interaction networks** (undirected collection of pair wise interactions on a set of proteins): Given a pair of weighted PPI networks (and a list of pair wise sequence similarities between proteins in the two networks) we need to find the best overall match between these networks.
- **Distributed and scalable solution for the existing biological databases**

Thank you!