Team

Jaroslaw Hirniak, s1143166

Michael Lotkowski, s1207068

Problem statement

WWW organised knowledge sematnics around pages and how they relate via hyperlinks. This idea gave rise to the Internet as we know it today and was used in all modern search engines, which give the pages the higer rank the more incoming edges they have.

However, this approach have drawbacks, mainly that the richer get richer and the poorer get poorer or in our version the popular get more popular and unpopular become even less popular. It is a fact that few websites get most of the trafic on the Internet -- Facebook has more than 1 billion of visits a day, most blogs are hosted on wordpress.com, media.com, or similar blogging platform, people use to media hubs such for sharing and accessing information, we buy everything on eBay or Amazon, access all articles on the Wikipedia, and in the result the web keeps converging to a single dot, what is not a desired effect. The web was made to connect and not to mash everything together into a single web page.

Importance

It is important problem as it often leads to very good sources of informationsuch as homepages, startups, insitutions, and interesting knowledge to remain unexplored on the Internet. The problem with the search as it is, it looks for popularity, and not for the utility. Hence even though no information exists in isolation, it can go as unpopular kid through the high school -- unnoticed.

We want to perform Milgram experiment and see how connected is knowledge, then compute graph metrics such as clustering coefficient, looking for communities, bridges, hubs, connectiviness, and structure (e.g. directionality of the information flow), maximum distance between nodes, superhubs, etc., and use that information to answer questions of how can we access the data (or aid user in doing so) efficiently, how to store it for quick access, and what else can we use the general knowledge about data for (e.g. can we use it to generate books?).

Dataset

The Wikipedia is the best source for analysing the knowledge network as it is available in whole for download and there exist representations used in previous research.

- Wikiseedia used in Wiki racer experiment, with shortest distance calculations for 4,106 articles: https://snap.stanford.edu/data/wikispeedia.html for experiment if the Wikipedia is small-world

network/
- XML dump of all articles on Wikipedia in English (14GB) - http://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2

Related work

How to address the problem

We will perform Milgram experiment on the knowledge base and measure mean/average/max distance between articles in terms of hops between them. Any such experiment could not be performed on the whole set, so sampling subnets from it to get significant results, finding the extreme cases, and mapping the structure of it, is going to pose interesting challenges in itself.

We expected mean and average distance to be 3 (based on experiments up to date), clustering coefficient to be high, homophily in topic domains, some superhubs such as geographical areas, core concepts, and common things to many fields (e.g. water), high expansion.

We will look into plotting heat maps based on how many times node is used in shortest path traversal, plotting histograms by article distance (to all other articles, with specific property, etc.), in/out edges, of certain properties, etc., statistics, influencial nodes, to how many articles the article is connected, and will compare it to random graphs.

We will use Python and IPython notebook with Gephi, NetworkX, Numpy, and similar tools.

After obtaining the measures and understanding how articles on the Wikipedia are related, we will think about how this knowledge can be used in order to * improve search algorithms by propagating page endorsement and ranking page by utility and less by popularity; * predicting what article the user is looking for and teleporting him to it; * predicting what article/product (ad) the user would be interested in based on the article trail and suggesting it; * characterising the user based on the article trail; * coming up with a novel way of distributing knowledge-base based on what articles should be hosted together based on the expected trails for use in NoSQL databases, most notably graph-based; * navigating the information search in a network where we only have local information about the graph (what is best strategy? greedy like in routing across the Internet?); * investigating what are efficient ways of traversing information to find what we look for, in learning, etc., and comparing it to information scent, information foraging (like animals forage for food), orienteering, berrypicking, etc.; * efficient system aiding in the article navigation (improving the experience); * how is number of hops per article per user related to education, experience in the area, highest education level achieved? and how this information can be used to help provide information at the most appropriate level to accelerate learning process; * cut-search - find article in the work cross-section and not described precisely by words; * book generation based on graph structure and findings; * predicting the next article, idea, etc. by observing changes in the network structure (historical data); * can existing network structure be used in creating argumentations? * and last, but not least, optimal wiki racer, without any special knowledge of the graph structure (e.g.

just network model).

Reference list

1. (2012) Human Wayfinding in Information Networks. Robert West and Jure Leskovec. WWW 2012, Web User Behavioral Analysis and Modeling
2. [(2004) The perfect search engine is not enough: A study of orienteering behavior in directed search. J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. In CHI, 2004.] (http://haystack.csail.mit.edu/papers/chi2004-perfectse.pdf)
3. (2009) Wikispeedia: An Online Game for Inferring Semantic Distances between Concepts. Robert West, Joelle Pineau, and Doina Precup. Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09).
4. Wikispeedia - game based on finding the shortest path thorugh Wikipedia articles from article A to article B.
5. Wikipedia: Book creator tool.

Data sets list

1. Wikiseedia with shortest distances:
    1. Paths and graph
    2. Articls in plaintext
    3. Articles in HTML
2. XML dump of all articles on Wikipedia in English (14GB)