CS224W – Reaction paper (October 18, 2012)

Information Network Search – Why do humans abandon Wayfinding?

Aju T. Scaria (ajuts), Rohan Kamath (rdkamath) and Rose Marie Philip (rosep)

1. Introduction

Searching and trying to connect bits of information is a task humans do on a daily basis, whether it be in information networks like cross-referenced dictionaries or citations, or in social networks. In this paper we explore the different traits of human navigation through information networks, the strategies that help them find the information sources that they need and finally try to understand why some of the searches end up abandoned. We believe, finding an answer to this question will help us to organize information networks in a better way and to predict whether someone will quit a search before he/she reaches the target based on the first few steps. This would enable us to attract and retain more users/customers in websites either by providing more intuitive and navigable link structure or by offering assistance to a few people who we think have a higher chance of getting lost.

This paper is built in the following format. First, we give the summary of three research papers we found interesting. This is followed by a critique of each paper and brainstorming for ideas and possible extensions that could be done to better understand why people abandon searches. This helps us formulate the project proposal (also including details of the dataset we will be using) in the next section. We conclude with references to the papers studied.

2. Literature review with summary

We reviewed the following papers in the area of information network search:

2.1. Human Wayfinding in Information Networks [Robert West, Jure Leskovec]

This paper studies goal-directed human search paths and identifies strategies that people use when navigating information spaces. Using this information, a model is also built to predict the target a user is trying to approach based on their first few clicks.

The data for this study was obtained from wikispeedia.net which is a web search game based on Wikipedia. Given a start and target article on Wikipedia, the user has to navigate from start to target only by clicking links encountered along the way. The highlights of this paper are:

- Search is easier while far away and close to the target as humans navigate through high-degree hubs in the early phase, and then once they reach near the target, the search is guided by topic/content similarity.
- Effective paths taken by humans are not much longer than the shortest possible path between the source and target articles.
- People who choose better hubs find better paths.

2.2. Automatic Versus Human Navigation in Information Networks [Robert West, Jure Leskovec]

The paper assumes that humans have a lot of background information about the structure of networks like Wikipedia, which they use to navigate the network. The paper tries to find the

importance of this structured knowledge and reasoning skills in path finding. The data for the experiment was again obtained from wikispeedia.net.

A basic navigation (search) algorithm was created that looked at all potential neighbors picked by an evaluation function. Heuristic agents experimented for this evaluation function was Degree based navigation (DBN), Similarity based navigation (SBN) and Expected-value navigation (EVN). Supervised learning and Reinforcement learning algorithms were evaluated using different features, which in conclusion performed better than heuristics. The key highlights of this paper are:

- A simple agent with basic knowledge of immediate neighbors and target can outperform humans, undermining the importance of the skillset and general knowledge of humans about how the topics are related.
- Agents perform better than humans on an average, but humans are less likely to get totally lost.
- Feature analysis showed that degree should be weighted less strongly on later parts of the search and features capturing similarity between next article and target article dominates on all positions especially as path progresses.

2.3. An Experimental Study of Search in Global Social Networks [P. S. Dodds, R. Muhamad, D. J. Watts]

The paper studies an experiment in which e-mail users attempted to reach one of 18 target persons in 13 countries by forwarding messages to acquaintances. But it is important to note that this study was done on social networks and hence, the results could be different from what is observed in an information network. Some of the interesting results found in this study were:

- There was more number of weaker ties (friends of friends) in successful chains than in unsuccessful chains.
- Senders decided the next recipient based more on geographical proximity of the acquaintance to target and similarity of occupation than the number of friends they had.
- Presence of highly connected individuals (hubs) appeared to have limited relevance.
- No evidence of message funneling.
- Supports random failure hypothesis and proposes a correcting component for attrition.
- Paper concludes that in order to do meaningful analysis of network, the strategies and actions of users need to be known.

2.4. Relation of topic with class

The class lectures cover Milgrams experiment (in which letters were sent to a stock broker by passing through friends) and Columbia small-world study [3]. The drop-out rates in these experiments were quite large (Around 65% dropout per step and 1.5% completion rate in Columbia small-world study). Even though the drop-out rates were quite similar at different path lengths, we believe multiple reasons would have resulted in the observed pattern. In [3], each of the individuals in the path had to decide whether to continue the path or not. This supports the random failure hypothesis. We question this hypothesis when it comes to Wikispeedia game where one person is responsible for the entire search path.

3. Critique and brainstorming for project ideas

3.1. Human Wayfinding in Information Networks

1. The paper uncovers many relations between network properties and its influence on finding target articles efficiently. The plot of evolution of article properties along search paths shows different behavioral patterns at different path lengths. We plan to use similar analysis on more network properties and content similarity of articles to understand the reasons for the high drop-out rates observed at different stages of the game.
2. The target prediction mechanism developed in this paper is quite effective in finding the possible targets based on the first few clicks. We would consider the feasibility of developing an algorithm based on which we could predict if a search is going to be abandoned by observing the first few clicks.
3. The paper establishes that Wikipedia articles conform to rank-based friendship model. This demonstrates why humans can navigate between articles in a complex network using short paths, even without knowledge of the global structure of network.
4. Even though around 54% of the searches in the game were abandoned, the paper does not analyze the abandoned paths and their properties. The paper assumes that since the drop-out rate is constant over the different path positions (around 10%), it might be a random phenomenon. But, we feel that different reasons might dictate people quitting at different stages. Some possible reasons could be:
    a. If a person quits after clicking one link, it might be because he did not like the game, got bored of it or he did not understand how it works.
    b. If the person quits after two clicks, it is possible that he navigated to a hub, but he couldn't easily find the next article to go to which would take him closer to the target article.
    c. If he quits at a later stage, it is possible that he started off from a hub in some direction, but he found that the topic of articles in the path is completely different from the target.
    d. It is possible that the target itself was very hard to spot in the network because it was obscured away from the main connected components as it had a very few in-links.
5. The paper does not give a lot of emphasis on back-clicks. The metrics to measure the efficiency of human search does not include back-clicks. We feel that there is higher chance of a search getting abandoned when more number of back-clicks are involved, as it is an indication that the user is finding it hard to navigate to the target.
6. The paper conjectures that the effective path found by humans are not longer than 2 clicks as compared to the shortest possible length (SPL). But, this is when we compare the SPL with effective path length of users, which does not consider back-clicks and does not consider drop-out correction. Once both of these are factored in, the median and mean grows to double and triple respectively, which is a very significant increase.
7. The paper assumes that people target hubs in the first few steps and then slowly directs the search towards the goal. But, since there are a lot of hubs that are reachable from a start article, choosing one of them over the other plays a key role in the effectiveness of the search. Especially, once we click an out-link from one of the hubs that wasn't the optimal

one, the search might stray away from the shortest path. Hence, the users need to have a global direction in which the search should proceed. This is not covered in this paper, but [2] takes care of addressing this.

8. The paper makes the conclusion that people who choose better hubs find better paths. 'Best hubs' would ideally be the ones with the highest degree. But in Figure 7 in [1], the average hub quality is only around 0.55 (Hub quality is defined as the degree of the second article, divided by the degree of the maximum-degree neighbor of the start article) and hubs in optimal paths have been found to have a degree 20 lesser than the ones humans chose, which does not provide enough evidence for the claim that better hubs lead to shorter paths. So, the degree of the hub is not the only criteria in choosing a hub to navigate to, otherwise, the hub quality would have been closer to 1. It is possible that the source and target articles may belong to entirely different topics, and people try to go to the nearest "switching-point" from one topic to the other, which just happens to be a hub. An interesting evaluation will be to measure the effectiveness of search paths in the Wikispeedia game that took a path via the hub with highest degree.

3.2. Automatic Versus Human Navigation in Information Networks

1. The paper brings to light the fact that unlike humans, automatic agents inspect all neighboring articles. This explains why humans often miss good opportunities since they have certain notions about the world and form a high level plan regarding the route before even making the first click.
2. The paper improves on some of the techniques used in [1]. It uses reinforcement learning to improve the results earlier achieved by logistic regression.
3. The paper assumes that humans have a pre-meditated plan which they execute to reach the target article. If they are not able to spot the links easily, they tend to abandon the search. But, there is no quantitative analysis if this is indeed true.
4. The paper suggests the major differences between automatic and human navigation – humans keep backup options while navigating to the target so that there are lesser chances to fail. As a result, they take longer paths but have more chances of success. This is corroborated by the metrics derived in [1] where it is mentioned that the lucrative degree of the hub is about 3.5 for human searches while that for optimal paths is 2.

3.3. An Experimental Study of Search in Global Social Networks

1. People need some incentive to continue the game. The incentive could be just the confidence that the target is reachable. This is supported by the fact that Target 5 received almost 44% of all completed chains in spite of his 'true' reachability not being very different from those of other targets. The senders allocated to him possibly had more belief that they would reach the goal because the target was a professor and 85% of senders were college educated. This leads us to consider the possibility that people might be abandoning the search in Wikispeedia when they feel the target is not easy to reach or is not 'popular' enough. This can be quantified by measuring the shortest path length or content similarity of the last few articles to target that the user visited before abandoning the search.

2. The paper has good metrics to support the random failure hypothesis. More than just the fact that a constant fraction of people left the game at each step, the survey conducted brings to light that only 0.3% people claimed to have left the game due to inability to find a suitable recipient.
3. The study collected information directly from people as to why they chose a path as opposed to guessing the reasons based on network structure. This analysis shows different results from the results in [1] where the analysis is based entirely on network structure.
4. The paper mentions that the presence of highly connected individuals (hubs) appears to have limited relevance to the experiment. But, there hasn't been any mention about who is considered as a hub, either in terms of reachability or the number of in-links or out-links.
5. The definition of the term reachability is not clear. It is somehow related to the path lengths of completed paths and not on the basis of network structure. So our hypothesis is that the more frequently visited nodes need not be the well-connected nodes according to network structure. This hypothesis could be checked by finding whether the ease with which a node can be accessed in an information network closely relates to its page rank. This can evaluate how empirical reachability compares to automated scoring algorithms like page ranks.

4. Project proposal

4.1. Project ideas and evaluation criteria (in parenthesis)

1. Evaluate how the path properties varied between searches that were completed successfully versus the searches that were abandoned. What are the main differences of successful and unsuccessful searches? Did the searches end because of a first few bad steps? (Use degree of nodes along the path or difference in similarity of articles with target article along the successful and unsuccessful searches.)
2. Evaluate the influence of the number of back clicks on search efficiency. (Do more back clicks mean more chance to abandon?)
3. Categorize different reasons why people might have abandoned a search. Some reasons could be:
    a. Bored / did not know how game worked (Path length <2?)
    b. Took a few wrong steps and got lost in the network (Path length > 2, similarity of last few nodes very less as compared to target?)
    c. Target too hard to spot in the network (Target has very few in-links/ page rank / number of people who completed a search to that article?)
4. Predict if a search will proceed to completion based on the first 'k' links clicked (Use machine learning algorithm and evaluate the prediction success on held-out test set).
5. Is there a direct relation between page rank of an article and how easy it is to navigate to that article in the game? Page rank is based on random walks on the web graph while the game is based on methodical navigation. (Measure if two articles having similar page ranks are visited with same frequency in the network).
6. What makes navigation difficult? How can web search be made easier? (Conclusion to the project)

4.2. Data

The data for the project based on wikispeedia.net was obtained from http://infolab.stanford.edu/~west1/manas/data/. We have a dataset which contains list of articles, their categories, links, page rank table, shortest path distance matrix, tf-idf similarity between each pair of articles, click trail of games (with and without back-clicks) and article adjacency matrix. Around 54% of the searches in Wikispeedia were abandoned. Since each game is played by one user throughout, there is higher chance of disinterested people not starting the game in the first place. Our hypothesis is that random failure might not be the major cause for attrition in a game like this. This provides a rich dataset to study the causes of abandoning search.

5. References

[1] Robert West, Jure Leskovec. Human Wayfinding in Information Networks: *WWW 2012 – Session: Web User Behavioral Analysis and Modeling.*

[2] Robert West, Jure Leskovec. Automatic Versus Human Navigation in Information Networks: *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media.*

[3] P. S. Dodds, R. Muhamad, D. J. Watts. An Experimental Study of Search in Global Social Networks. *Science 301(2003), 827*