Analyzing Wikispeedia Navigation Path

Jeong Hwan Kim and **Seong Min Yeon**

Computer Science & Engineering

Seoul Nat'l University

Abstract

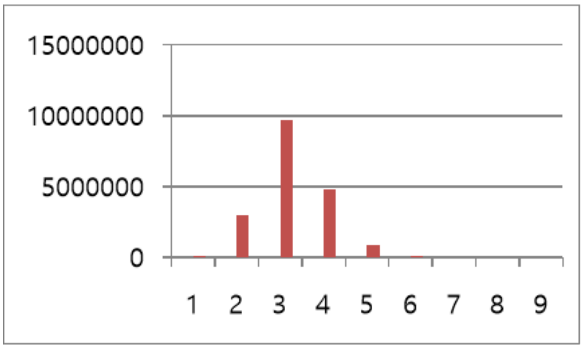
'Wikispeedia' is the human-computation game. In wikispeedia, users are asked to navigate from a given source to a given target article, by only clicking Wikipedia links. We will calculate shortest path and average path between two words in game, find out local search method and compute semantic distance between words by getting data from game.

1 Introduction

Common-sense knowledge is important for intelligent computer applications. For instance, we know a wheel is a component of a car. But it is hard to make computer know relation between them. There have been some research about wikispeedia. In that research, they calculate semantic distance by their own methods, analyze about wikispeedia network structure. We modify their methods and make our own method of evaluating semantic distances.

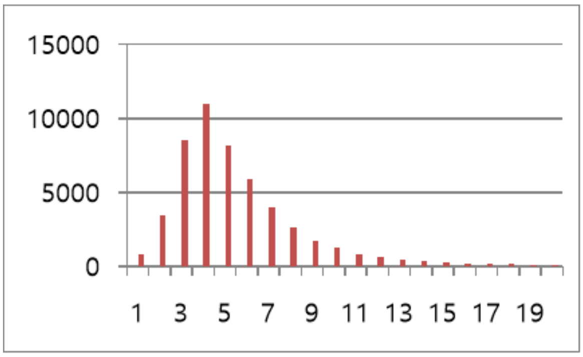
2 Network and finished path analysis

In wikispeedia game we can't know global structure of the network. So they use their local search method for game. In contrast, we can know global structure of wikispeedia for network analysis. We can know optimal path of the game for any start node and destination node by using Floyd Warshall algorithm and the network.

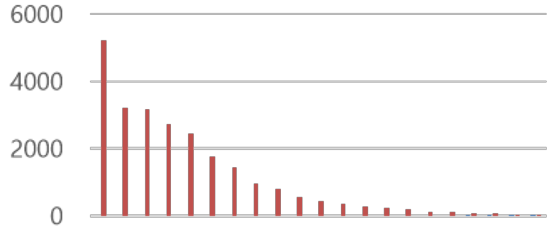


Above graph shows length of optimal path. Number of articles is 4604. And possible connected node pairs are 18588235. The average path length of optimal path is 3.2024 and all optimal path have less than length 9.

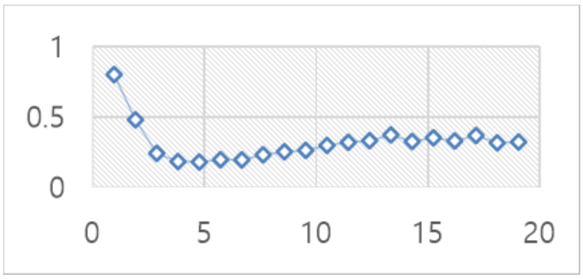
Next, we calculate the average path length of finished paths. Number of finished paths is 51318. The average path length of finished paths is 5.75. Six degrees of separation can be applied to this result. We can guess this network has similar attribute with real network. Below Graph shows the number of finished paths depending on path length. Compared to optimal path, finished path have long tail.



In this wikispeedia game, number of finished path is 51318 and unfinished path is 24874. There are too many unfinished path, so we have to check about drop rate and use it for semantic distance later.



Above graph shows the number of unfinished path depending on path length. Different to finished path there are many unfinished path when path length is 0 and monotonically decrease when path length increase. Using above 2 graphs, we can get drop rate.

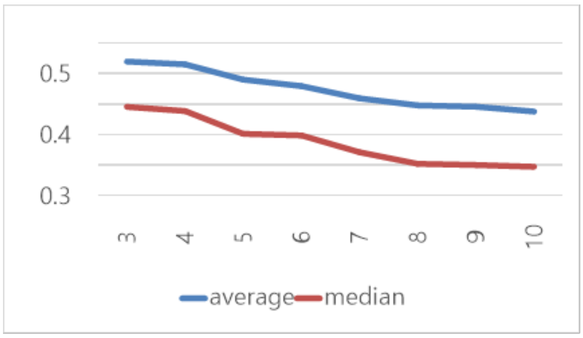


As we can see, rate increase after path length is 5. When path length is less than 5, drop rate looks very strange. We decide to ignore that part because when length is 0 or 1, there are too many games unfinished because of timeout, which means that no click was made for 30 minutes. We think this kind of results should be modified, so in this report, we only think about when path length is more than 5 and less than 15. Because the game with length more than 15 has a few result for analyzing. In that range, we can get drop rate follow 0.0205x + 0.1189. We will use it for semantic distance later.

3 Find out local search method

We learn six degrees of separation concept, which posits that any two people on Earth are six or fewer acquaintance links apart. For example, kevin bacon number shows six degrees of separation concept in movie actors and movies. There are also similar experiments by Milgram. He asked people to forward a letter to a stock-broker in Boston by passing it to their friends and then measure the number of forwards. The result of this experiment shows the average chain length is 6.2, we can see six degrees of separation in real network. In the experiment, people sent letters by geographic and occupation methods. We show six degrees of separation can be applied to wikispeedia navigation path. We want to know the local search method in wikispeedia game. We hypothesis that people will choose article which has many links at the beginning of the game and they select similar articles at the end of the game. People select the hub, article which has many links to other articles, so they can move to any concept of articles. And then at the end of the game, they select similar and more specific and concrete words so they can approach to the destination.

There are 51318 finished paths. We have to calculate start article's neighbors out degree and similarity between last article and before last article, So we only calculate about paths which have length 3 or more. After this data modification, 47085 paths left.



First, we calculate the average out degree of start article's neighbors and the out degree of the second article in finished path of the game. Average of the average out degree of start article's neighbors is 53.81494 and the average of the second article chosen by people in finished path of the game is 85.63188. We can see people choose article which have more out degree than the average one. For more concrete proof, we check relation between path length and the ratio about out degree, second article's out degree chosen by people per max out degree of first article's neighbors. We can see average value and median value of the ratio monotonically decrease when path length increase 3 to 10. By above results, we can say people choose the article which has many links for the first of the game.

Next, we check the strategy about the end of the game. We check similarity by 2 methods. We check similarity by coefficient and adamic/adar score between 2 articles. Our destination article and before that article chosen by people has coefficient value 0.111198. Before last article and before that article have coefficient value 0.098944. And the average coefficient value of all article pairs is 0.052791. In this result we can see that people choose similar articles at the end of the game. By progressing above research with adamic/adar score, last 2 element's coefficient value is 3.819316 and before that has 3.90233 and the average coefficient value for all node pair is 3.042493. Adamic/adar score also show similar articles are selected at the end of the game.

4 Clustering graph

Using snap api we cluster graph by CNM algorithm and Infomap algorithm. Using CNM algorithm, graph divided into 9 communities and modularity result is 0.296486. Using Infomap algorithm average code length come out as 2006 and the number of communities is 295. For using above algorithms, we change directed graph to undirected graph and progress algorithms. There are also clique percolation algorithm and Girvan newman Method in snap api. We try it but can't get the result. Results of using clustering methods provided by snap api is unsatisfiable, so we try to get semantic distance for comparing two words instead of clustering network.

5 Semantic Distances

In the previous research, they make semantic distance by assuming that uniform prior click probability. So they make a calculus as below.

$$p^{0}\left(\mathrm{A}^{'}=a^{'} \,\middle|\, \mathrm{A}=a, \mathrm{G}=g \right)=1/\mathrm{L}_{a} \quad (1)$$

A, A' and G be random variables representing the current Wikipedia page, the next Wikipedia page and the goal page of a game. And L_{a}means a's out degree. Using probability made by above calculus and pageRank, they made their own semantic distances. They consider one particular path p which reach goal through a1, a2, … , an.

$$d_{p}\left(a_{i}, g \right) = \frac{-\sum_{j=1}^{n-1} log\mathrm{P}^{*}(\mathrm{A}^{'}=a_{j+1}|\mathrm{A}=a_{j}, \mathrm{G}=g)}{-log\mathrm{P}age\mathrm{R}ank(g)} \quad (2)$$

For getting a path-independent distance from a to g, they get the average. If there are m paths running through a and reaching goal g. Below calculus shows the average value.

$$d\left(a, g \right) = \frac{1}{m}\sum_{k=1}^{m} d_{p_{k}}\left(a, g \right) \quad (3)$$

As we can see from the part2, there are too many unfinished path in the game. So we think we have to apply that result for analyzing the data. So we will use drop rate getting from data. We can modify click probability using drop rate as below.

$$p^{0}\left(\mathrm{A}^{'}=a^{'} \,\middle|\, \mathrm{A}=a, \mathrm{G}=g \right) = \left(1\text{-}d\left(r \right) \right) *(\frac{1}{\mathrm{L}_{a}}) \quad (4)$$

d(r) means drop rate at that status, we get it as 0.0205x + 0.1189. For instance, at prior click probability, drop rate becomes 0.1394.

6 Further plan

We will use our semantic distances for a particular words and list words by semantic distances. We will choose a word which was chosen by people most in the game. We think the more data number, the better and exact result we can get it from. Getting related words by above process, we will compare our result with other algorithms by showing people and asking to select words they considered most closely to the target.

7 References

Robert West, Joelle Pineau, and Doina Precup. Wikispeedia: An Online Game for Inferring Semantic Distances between Concepts. In *21st International Joint Conference on Artificial Intelligence (IJCAI'09),* pp. 1598–1603, Pasadena, Calif., 2009.

Robert West and Jure Leskovec. Human Wayfinding in Information Networks. In *21st International World Wide Web Conference (WWW'12),* pp. 619–628, Lyon, France, 2012.

Robert West and Jure Leskovec. Automatic versus Human Navigation in Information Networks. In *6th International AAAI Conference on Weblogs and Social Media (ICWSM'12)*, pp. 362–369, Dublin, Ireland, 2012.