Six by Two Degrees of Wikipedia using BFS Algorithm

K Jitendran 1, S Akash 1, N Aswin 1 and T Rajesh Kumar 1
1Department of Information Technology, Sri Krishna College of Technology, Coimbatore, India.
jitendran250@gmail.com, akash.s1710@gmail.com, aswinneson7@gmail.com, rajeshkumar.t@skct.edu.in

Abstract. SixbyTwo Degrees of Wikipedia is a concept that is built on the idea of Six Degrees of Separation, where articles can be connected within six steps. This theory motivated the invention of online communities, after succeeding in the real world. The above concept has been visualized for connecting two Wikipedia articles with a minimum number of articles connecting them, by using a modified BFS algorithm. This paper is helpful for students and researchers, for their research. The graph approach formulates the theory and evaluates the minimum number of links between any two given articles. This approach provides much faster results than other algorithms and the traditional BFS algorithm.

Keywords: BFS, Six Degrees of Separation, Graph Algorithm

1. Introduction
The popular hypothesis, Six Degrees of Separation, holds that any article or people can be connected by a chain of no more than six links. Characters of Wikipedia can be scrutinized to see whether a similar ideology holds for its articles, SixbyTwo Degrees of Wikipedia aims to be a collection of the following things: The articles separated by the prolonged minimal chains in the encyclopedia (especially where there are more than three links). Generalization over time, where possible, seems to produce the precise link between articles. Failing this, generalization is tried over space to reduce the chain length. More intuitively remote articles are separated by precise chains — chains of fewer than three articles. That is, there should be articles that have no obvious connection but which, however, can be linked quickly in Wikipedia. These subjective categories should still be interesting here, it may be sufficient to describe them “interestingly" or "curiously" precise chains. Chains with links between two articles and people are selected at random.

2. Overview of the system
SixbyTwo degrees of Wikipedia has the motive to find the precise path between articles, preferably under three connections away from each other. As a final result, a set of "an article of an article" statements can be used to relate any articles and people within three steps. The modified BFS algorithm is useful in making faster connections between the given two articles.
3. Proposed system

SixByTwo Degrees of Wikipedia checks whether an article can be reached from a source article, within three clicks. It opens the Wikipedia page of the given article and then parses the HTML page, to find the hyperlinks present in the current article, that redirects to other articles. After finding the path between the articles, they are opened as separate tabs in the Google Chrome-browser Figure 1.

![Figure 1. Work-flow Diagram](image)

4. Breadth-First search algorithm

A modified version of the Breadth-First Search algorithm to find the hyperlinked path between articles. Like Dijkstra’s, Breadth-First Search and Depth-First search precise path is also a solved problem, Yan ZHOU [1-3]. One of the solutions to find the precise link between any articles or people can be found with the help of Breadth-First search algorithm. It is an algorithm which performs searching in a wide range by looking at each node pre order i.e., looking a node and then looking at all the nodes adjacent to the node first [4-9]. In breadth first search method, level n nodes will be looked first and level n+1 node will be visited next. The root node is searched first from left to right and moves to 1st Level and 2nd level and etc. [10- 15]. The main profit of using the BFS algorithm is that it will not stop or get stuck at any point. If it has more than one solution the Breadth-First search algorithm will find all the possible keys and will display the solution which has the minimum number of links as the output. The algorithm has a weakness, that it requires a large amount of storage, because it keeps all the nodes during execution as a list and requires a bit extra time to
run and test the nodes on level \( n \) before reaching the articles on level \( (n + 1) \). But this algorithm is the most suitable and stable algorithm when compared to other graph algorithms. The Depth-First Search algorithm cannot be used for this case, because the articles are parsed first in a fashion that causes the search to proceed infinitely. Thus, this makes BFS the best algorithm for the task at hand. Breadth-First Search algorithm develops root nodes, and then formulating all feasible moves, thereby producing fully developed conditions without violating the terms. When no new nodes are present or when the connection reaches a certain threshold the set, in this case more than 3 articles or when other conditions could be revealed. The illustration of the BFS process can be seen in the subsequent Tree process. The articles are processed in such fashion until the discovery of the solution Figure 2.

![Diagram](image)

**Figure 2.** Breadth First Search procedure

5. Modules

5.1. Article Name

The module is used to modify the name of the article such that it can be used as a part of the Wikipedia URL. The spaces are replaced with underscores (_), and the words are always capitalized. This is then processed to store in a local variable. This variable is now used in various places throughout the program. This then gets back to the calling module.

The module then fetches the status code of the article. The article name is converted to an URL and then passed to a function where the article is validated. If the article is present it will return a response of ‘200’, else it will return a ‘404’ status code meaning that the page is not present. By this, the user input for the article name is modified and checked whether it points to a valid Wikipedia article

5.2. Extraction Link

This module is used to get the URL for the given article. The user input for the source and the destination article is converted to a URL in this module. In addition to that the module passes the obtained link and obtains all the hyperlinks present in the article page. This is done by using the XML parser in Beautiful Soup Library. The module opens the given article and parses the html page and obtains the link in the page, refines them and adds them to a list along with the information of the current level and its parent article. Potentially unwanted links such as images, disambiguation articles and other help articles are generally avoided with the help of this module.

5.3. Find Link

The module is the core of the paper. It is where the BFS algorithm is executed. First boundary conditions like whether the source and the destination article are the same, is checked. Only after this check the
algorithm is executed. The algorithm uses a queue data structure here in order to implement the BFS algorithm. The path between two articles is found using an alternate version of the BFS Algorithm. We use a visited set here in order to avoid parsing links or pages that have already been visited. This modification is very useful as this avoids unnecessary computation and saves memory. This also increases the speed of the algorithm significantly. While parsing the algorithm it tracks the number of pages parsed. It parses an article and stores them in a list and a set. Then the links are searched even before storing them in the list and set thus making the search even more faster as the number of links parsed and stored is significantly reduced. The path between the articles is found and then they are passed to the new tab module.

5.4. New Tab
The module is used to open all the connecting articles between the source and the destination article, inclusive of both articles respectively, as a new tab in Google Chrome-browser. The module employs the Selenium web driver to achieve this. This is the final level of output displayed to the user, in addition to the metrics shown in the Find Link module.

The articles are stored in a stack and then they are opened one after the other, separately as new tabs in the Google chrome web browser.

5.5. Main Module
It is the main connecting module where all the above-mentioned modules are connected. This module powers the entire operation and is the brain of the program. The start and the end articles are obtained from the user. A call is made to the functions in the article name module in order to refactor the names and process it. Then, a call is made to the Find Link module with the start and end article names as arguments, where further processing takes place. The module is responsible for timing the entire runtime of the algorithm.

6. Experimental result
The search will run for two articles that are entered by the user. The articles are then searched and the connecting articles, including the source and the destination articles are then printed in the console. In addition to that, the time taken for executing the algorithm is printed along with the number of articles that have been parsed in order to find the connection between the two given articles. Then the tabs are opened in the order of the articles found.

In this particular case, the source article is ‘Klingensmith’ and the article to be found is ‘Joe Biden’. The articles are entered and the names are refactored for the URL. The connection exists between the two articles. The articles can be connected with a minimum of 3 steps. The tabs are then opened in a separate browser window Figure 3, Figure 4 and Figure 5.

The articles in order are:
Klingensmith -> John Klingensmith Jr. -> United States House of Representatives -> Joe Biden.
This is the most effective way to reach the destination article from the source of origin article.
The program executed in 4.6 seconds.
The number of links parsed is 2997 articles.
7. Conclusion and future enchantments
The project, thus, provides a better solution for students and researchers by easily identifying the relationship and links between the articles within three clicks. SixbyTwo Degrees of Wikipedia is now efficient in analysing the link and finding the probability of similar articles. The time consumed in processing the link for the article can be reduced. In future the system can be elevated by increasing the accuracy.
References

[1] The Small World Problem, http://en.wikipedia.org/wiki/Small_world_experiment.

[2] Duncan Watt, Peter Sheridan Dodd E. J. Newman, Identity and Search in Social Network, 2002 Science, Vol. 296, no. 5571, pp. 1302 – 1305

[3] Z. Zhou, U. Wang, D. He, and Z. Wang, A fewest-turn-and-shortest path algorithm based on breadth first search, Geo-Spatial Inf. Sci., vol. 17, no. 4, pp. 201–207, 2014.

[4] I. Silvela and J. Portillo, Breadth-first search and its application to image processing problems, IEEE Trans. Image Process., vol. 10, no. 8, pp. 1194–1199, 2001.

[5] Rajesh Kumar Thangavelu, et al, Blackspot Alert and Accident Prevention System 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE Xplore, 10.1109/ICCCNT45670.2019.8944412

[6] Robbi Rahim1, Dahlan Abdullah2, Saiful Nurarif3, Mukhlis Ramadhan3, Badrul Anwar3, Muhammad Dahria4, Surya Darma Nasution5, Tengku Mohd Diansyah6 and Mufida Khairani6, Breadth First Search Approach for Shortest Path Solution in Cartesian Area. 2017.

[7] U. Agarwal, F. Petrini, P. Pasetto, and D. Bader, Scalable Graph Exploration on Multicore Processor, in Supercomputing. Proceedings of the ACM/IEEE SC 2010 Conference.

[8] Nicholas Pippenger. On simultaneous resource bound (preliminary version). In Proc. 20th IEEE Symposium on Foundations of Computer Science, pages 307-311, 1979.

[9] Rajesh kumar,T Samsudeen, S, Sangeetha, S, Sudha Rani, Enhanced Visual Attendance System by Face Recognition USING K-Nearset Neighbor Algorithm Journal of Advanced Research in Dynamical & Control Systems, Vol. 11, 06-Special Issue, 201

[10] S.H Yen, D.H. Du and R. Ghanta, Efficient algorithms for extracting the K most critical paths in timing analysis, 26th ACM/IEEE Design Automation Conf., July 1989, pp. 649- 654.

[11] Klaus Schmidt and Ece H. Schmidt, A Longest-Path Problem for Evaluating the Worst- Case Packet Delay of Switched Ethernet, Department of Electronic and Communication Engineering, Cankaya University, Ankara, Turkey.

[12] Cormen, Thomas J., Leiserson, Charles E., and Rivest, Ronald M. 1992. Introductions to Algorithms. Cambridge, Massacchusetts, USA: McGraw-Hill Companies, Inc.

[13] Rajesh kumar,T., Remmiya devi,G., Abinaya,K., Deepika ,N N., Priyadharshini,S., An Integrated Density Based Traffic Load Balancing System in a Cloud Environment, Pak. J. Biotechnol. Vol. 14 (4) 623627 (2017), ISSN Print: 1812-1837, ISSN Online: 2312-7791

[14] Fahmid Al Farid. Md. Sharif Uddin. Shohag Barman. Amirhossein Ghods. Soumitra Das. Md. Mehedi Hasan. A novel approach towards a parallel implementation of Breadth-First Search algorithm using a graphic processor unit. 2014.

[15] Gjoka N, Kurant, C.Butts, and Markopoulou B, Walk in facebook: A case study of unbiased sampling of Osn, Infocomm, 2012.