Finding Related Pages to Beat the Wikipedia Game

**Introduction:**

Wikipedia holds a massive amount of web pages connected through a series of hyperlinks. This makes Wikipedia’s data the perfect candidate for graph theory analysis. We hope to be able to relate the massive number of pages.

With the ability to find the relations or “closeness” of two wiki pages, we may create a script that helps us play the Wikipedia Game!

The Wikipedia Game is a game known often played by school children in computer labs. The rules are very simple. First, two wiki pages are chosen at random. The player chooses a starting position and then attempts to navigate through only hyperlink clicks to the other wiki page.

For example, say the two random wiki pages are on Frogs and Airplanes. One solution would be to “hop” along hyperlinks as follows: Frog -> Madagascar -> World War II -> Jet Aircraft -> Aircraft -> Airplanes

The idea is that using our end goal, we may find similar pages within the hyperlinks of the starting page.

**Data Collection:**

Wikipedia has large data dumps available to the public. The English-only pages come to about 60G compressed, which is more than enough to play with. A smaller subset may be used initially.

**Equipment Needed:**

As far as I know, there should not be any additional hardware needs. No cameras or sensors would be needed as the data has already been gathered. As this problem could be thought of as a large depth first search graph problem, a true Hadoop distributed file system could be very helpful in computation.

**Language/Framework/Library:**

The current plan is to work with Java and Hadoop to take advantage of the MapReduce framework. It feels necessary due to the size of Wikipedia’s data.