Homogeneity measurements in parent-child webpage relationships
by Benjamin Livingood

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science
Montana State University

Abstract:
Since the advent of the World Wide Web, people have tried to codify its internal structure. Due to the increasing amounts of webpages being created, various web search engines have been created, each trying to serve the average user with large groups of interrelated webpages.

There are, howsoever, some problems with the webpages returned when a person makes a query to the search engine. Depending on the algorithms and metrics used by the search engines in categorization, a user generally finds that regardless how specific a search term is, inapplicable results are still returned.

With this in mind, the author proposes that a web search robot, basing decisions on a set of metrics pertaining to the quality of a page, would create a database that would have linked webpages that were as equally rich as its peers. In order to implement this, the author proposes the idea of homogeneity, or similarity between pages and the pages it links to. Basing decisions on trends in homogeneity, the search engine can then pare out pages that trend towards poor scores and thereby retain a certain assured quality.

Through the use and examination of homogeneity, search engines could prune the number of sites they were going to follow, initially or later, when building their databases. By pruning poor content sites from the search tree they could thereby increase the speed at which the databases were created while retaining a measure of quality.

If homogeneity holds true over generations, the benefits to search engines would be two-fold. It would achieve more relevant and faster generated databases. The databases would then be of more use to the average user in searching for certain topics.

HOMOGENEITY MEASUREMENTS IN PARENT-CHILD WEBPAGE RELATIONSHIPS

By

Benjamin Livingood

A thesis submitted in partial fulfillment
of the requirements for the degree

of

Master of Science

in

Computer Science

MONTANA STATE UNIVERSITY
Bozeman, Montana

May 2002

APPROVAL

of a thesis submitted by

Benjamin Livingood

This thesis has been read by each member of the thesis committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the College of Graduate Studies.

Dr J. Denbigh Starkey J Denbigh Starkey 5/3/02
(Signature) Date

Approved for the Department of Computer Science

Dr. J. Denbigh Starkey J Denbigh Starkey 5/3/02
(Signature) Date

Approved for the College of Graduate Studies

Dr. Bruce McLeod Bruce R. McLeod 5-3-02
(Signature) Date

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a master's degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library.

If I have indicated my intention to copyright this thesis by including a copyright notice page, copying is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for permission for extended quotation from or reproduction of this thesis (paper) in whole or in parts may be granted only by the copyright holder.

Signature 

Date 5/3/02

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

ABSTRACT

Since the advent of the World Wide Web, people have tried to codify its internal structure. Due to the increasing amounts of webpages being created, various web search engines have been created, each trying to serve the average user with large groups of interrelated webpages.

There are, howsoever, some problems with the webpages returned when a person makes a query to the search engine. Depending on the algorithms and metrics used by the search engines in categorization, a user generally finds that regardless how specific a search term is, inapplicable results are still returned.

With this in mind, the author proposes that a web search robot, basing decisions on a set of metrics pertaining to the quality of a page, would create a database that would have linked webpages that were as equally rich as its peers. In order to implement this, the author proposes the idea of homogeneity, or similarity between pages and the pages it links to. Basing decisions on trends in homogeneity, the search engine can then pare out pages that trend towards poor scores and thereby retain a certain assured quality.

Through the use and examination of homogeneity, search engines could prune the number of sites they were going to follow, initially or later, when building their databases. By pruning poor content sites from the search tree they could thereby increase the speed at which the databases were created while retaining a measure of quality.

If homogeneity holds true over generations, the benefits to search engines would be two-fold. It would achieve more relevant and faster generated databases. The databases would then be of more use to the average user in searching for certain topics.

STATE OF THE ART

Web search engines judge the pages indexed in a variety of ways from neighboring words to the frequency a particular word appears in the text of a webpage. Each page indexed through the use of web spiders, autonomous programs that read webpages, is then codified by various metrics that the search engine uses. On categorization, the search engines store the results into a continually growing database of grouped webpages.

To create a more relevant search mechanism, among using other scalar values, Google uses a mechanism called PageRank.[1] PageRank is a metric that for each page counts the number of pages that link to it. The PageRank of all indexed webpages forms, according to the authors, the principle eigenvector of the normalized link matrix of the Web. The overall idea for PageRank is that the more popular the page the more likely it is to have pages that link to it.[1, p.3] Therefore when a person enters in a search term in Google the more 'popular' sites get returned to the user.

Throughout the whitepaper upon which Google's technology rests, there are no claims to the quality of any indexed webpage. PageRank is seen as a form of self-regulation, where the users tend towards linking to more interesting and relevant webpages, thereby reducing the impact of information-poor pages on searches in general.

For this paper, we are going to look at the larger question, whether or not it is possible to form a measure based on a set of one or more metric for a webpage, thereby allowing for the culling of poor scoring webpages from searches. This takes then the need

for PageRank to be no longer a necessity, and more of a secondary ranking. Since calculation of PageRank can be time intensive, taking according to the authors a few hours for 26 million webpages, this calculation could be done more at leisure and thereby save computational time.

DATA COLLECTION

In the first phase of the research, the author decided to model the training data from a state of the art web search engine, www.google.com. In doing so it was hoped to collect already clustered data for comparison in the second and subsequent phases.

A precondition for choice of a clustered group was that the group produced copious amounts of data on that particular subject. The author chose webpages pertaining to the world of GreyHawk, due to their providing of information in large amounts on their hobby and thereby including a great deal of material in each webpage. The Montana State University Computer Science faculty webpages were also used to determine if any predictions made on the Greyhawk set were erroneous. The added benefit was that the authors of those pages could be contacted more readily in order to ascertain any anomalies that might crop up.

The second phase of the research then was used as an overall estimator. This was done in order to determine whether or not our sample size in the prior phase was a unique case or if some general principle was occurring.

METHODS

Phase I

The first step was to collect the links from the Google webpage after entering the keywords "Greyhawk AD&D". This was achieved, crudely, through the use of lynx and some hand entering of the proper URL query to Google's website. In doing so, the information return was free of hypertext markup and in a ready made form to then perform our tests.

Second, from the returned results of Google, we started extracting the links. This was also done through the use of lynx, which after "dumping" a website also dumps all the links that were in the target webpage.

The resulting links from the initial Google query were then used as a starting platform to gather our actual data. These links were first filtered for duplicates and then fed into a file to be used later to analyze and retrieve linked webpages.

Phase II

In phase II, a perl programming construct was used to crawl a website. This construct was created to do a depth first search following only a certain random subset of child links. At each step down the resulting tree of links, the current webpage was extracted into text, if possible, and was then stored in memory to be worked with after the search had reached a certain time limit.

Also, in phase II, the resulting search through the tree was recorded in the form of a hash with each parent node being a key value in the hash. Each subsequent key hashed to a unique list of the linked webpages, symbolizing the children in the tree.

ANALYSIS

Phase I

As expected with crude tools, the process was very time intensive as from 100 samples we had extracted 3513 linked webpages over the course of several working days. Since it was so time intensive, this made for an impractical implementation for large scale analysis. In addition there was always some overlap in the linked webpages.

The overlap tended to be rather minimal and was skewed towards certain main websites. These occurrences of overlap were overcome through the use of hashes, in which the 'parent' webpage was the key in the hash and the children where the values. Each addition of a new child link into the hash was checked versus the prior values and added only if it was unique.

Phase II

In phase II, having remedied the problems of overlap, the author noticed quite a drastic speed up in using a preconstructed Robot. By neglecting to delay while crawling a website, a practice generally frowned upon, 2396 starting webpages were gleaned in the course of a few hours.

In addition to a speed up, the robot was easily configured to take a random subsample from the starting parent webpages and thereby reduce the chance of erroneous entries or duplicate entries in the table from return trips to the webpages. For example, if

after visiting webpage Y and gathering a sample set T from the page a link in T referred back to Y, it was easier to get a new sample set S from the page.

FEATURE EXTRACTION

Once the data started pouring in from various places the question was what would be an accurate and easily implemented metric set for our features. In phase I, various English language metrics were looked at while for phase II only the entropy, or information, of a page was looked at.

Phase I

By using a Perl module called Lingua::EN::Fathom, which keeps various statistics on a piece of text, the author hoped to keep the duplication of code to a minimum during the analysis phase of the experiment. Upon extraction of data from the initial Google page, our starting 100 parent webpages gave us the set of metrics shown in Table 1:

Table 1. Sample Metrics from an Extraction Set

Number of characters	: 30760
Number of words	: 3698
Average syllables per word	: 1.76
Number of sentences	: 520
Average words per sentence	: 7.11
Number of text lines	: 708
Number of blank lines	: 125
Number of paragraphs	: 123
READABILITY INDICES	
Fog	: 9.24
Flesch	: 50.55
Flesch-Kincaid	: 7.98

Rather than predefine a predetermined set of classes, a set of continuous values were appraised. In doing so there would be no hidden classifier or set of classifiers that would cause the classes to perform just like the author expected.

In terms of established, semi-credible and continuous metrics, the author looked at: the Flesch Readability metric[2], the Gunning Fog index[3], the Salton Information metric[5], and, in the second phase of the study, the Shannon Average Information Index as outlined by Salton.[5]

Flesch Readability Metric

The Flesch readability metric was created to indicate the ease of reading a sample of text. It is based on a continuous scale from zero to 100, where 0 denotes being practically unreadable and 100 being easily read by anyone literate. [2, p.216] The Flesch readability metric is determined by taking 206.835 and subtracting the difference between 1.015 * the number words per sentence and 84.6 * syllables per word. [2, p.215-216]

Another system that Flesh detailed included a metric where one would take the number of personal words and personal sentences, and multiply those counts by 3.635 for the personal words and .314 for the personal sentences. Personal words and sentences are structures that contain person pronouns, names, or gender specific words. This is then purported to be a measurement of the human interest in a sample.

Flesch seems to advocate a usage of both the reading ease and human interest metrics in codifying samples of text, but since Lingua::EN::Fathom doesn't identify personal words overly well, this led to the use the Flesch reading ease metric.

Gunning Fog Index

The Gunning Fog index, as outlined by Robert Gunning, is calculated by taking the number words per sentence, adding the percentage of complex words in the sample,

and then multiplying that sum by 0.4.[3, p.40] The percentage of complex words is derived from the idea that any word with more than two syllables, that is non-hyphenated, is added to a running total over the sample and then divided by the total number of words. The result is then multiplied by 100 to get a percentage estimate of how "hard" the text is in relation to the amount of single or dual syllable words. According to Gunning, any sample scoring above a 12 in Fog index is harder to read than any normal publication.

Phase II

The Salton metrics were looked at in phase II in order to determine if there were any better metrics available. As per the Gunning, *et al.*, metrics, each of which rely on word and syllable counts, it was determined that any hypothesis based on these metrics would be reliant on the determination of constant values as put forth by their creators. These constants are intrinsically poor selections and can cause problems, since the authors had derived the constants based on "gut feelings" and some fair approximations of data they had in their time. Since that was over fifty years ago, and pertained to newspaper publishing, no concrete web data exists to prove or disprove their usefulness in this exercise.

As such, since Shannon's information theory does not take into account any constants and instead relies on frequency, it is intrinsically a more reliable classifier of all manner of data.

Salton/Shannon Metric

According to Salton, in the Text Analysis portion of his text, the Information Measurement of a selected text is based on the frequency of query words or sets of words in general inside each of the pertinent texts.

For example, if the word "the" shows up in the document once every 10 words its probability inside the text is 0.1 and its information measurement is detailed in Table 2.

Table 2. Example of Salton Information Metric

INFORMATION = -log (0.1) = -(-3.3223) = 3.223 [5, 64]

In addition, according to Shannon's formula, as outlined by Salton, we then take the negative sum of the probabilities of the query terms and the log of their probabilities in the document and thereby derive an Average Information Measurement per document with respect to the query terms.

In terms of this thesis, a pure information metric was used, in which the negative sum of the probabilities of the words multiplied by the natural log of their inverses was derived. It is a slight alteration of Salton's metric and proved to be more in line with Shannon's original paper on information theory.

FEATURE REDUCTION

With these features in mind, it was crucial, especially in phase I, to ensure that we were not recording additional metrics into our data that were in fact not buying us any additional information and merely were increasing the space in which we were working.

Phase I

In choosing to examine the Fog, Flesch, and Flesch-Kincaid forms of measuring a document's text exclusively the author initially passed over other attributes such as total words in a sample. Afterwards the author ran some correlation statistics on the samples to determine uniqueness of the variables. The overall idea behind this was to minimize the total number of factors that were used without sacrificing coverage. The correlation statistics are stated in Table 3.

Table 3. Correlation of Various Metrics

	Flesch Index	Flesh-Kincaid
Fog Index	-0.281188362416199	0.988245023338575
Flesh Index		-0.264397342200341

As one can see, the Fog versus Flesch-Kincaid is highly correlated, due in part to their reliance on the syllable per sentence count and also because they are supposed to roughly show the grade level required to read a piece of text without any trouble. Since the Fog and Flesch-Kincaid were so correlated the Flesch-Kincaid index was dropped. Further tests on various subsets of the data to determine correlation tended to flip-flop the correlation statistics between Flesch-Kincaid and Fog.

Now given two features for testing the author begin to get the data from each page's linked articles in turn keeping the Fog and Flesch metrics for each article in an overall file that corresponded to the main page. So for the 11th sample on the Google search the author had an 11th body statistics file and an 11th links statistics file. Then, in keeping with the hypothesis of similarity between content of main page and linked page content, the author then averaged the links' statistics and tossed them into a Super file that had both the body's statistics and now the link-average statistics.

Looking over the total population of Roleplayers and there was discovered the set of statistics (Table 4):

Table 4. Statistics of Sample Roleplayer Population

	Fog Metric	Flesch Metric
Mean	11.4316064257028	41.5775207496656
Standard Deviation	9.83406351330864	31.6355991970666

With such high variability in the means, the author knew that a scatter plot of the data could have a high probability of very little clustering. In addition to this, the files that contained the main pages' statistics and the linked pages statistics, if plotted, would push the analysis into 3-space (if not 4-space if special care was not taken as to how the data was plotted). Any increasing of dimensionality would further limit the ability of how the derived data approximated the "true" distribution of the data. Thus, any further examination of data would cause the analysis to be more problematic and circumspect.

Spatial Reduction

With the prospect of an ever-increasing space over which our test data was scattered, it was decided to re-examine the goals and hypothesis in the exercise. Since we were attempting to find a set of correlations between the content between main pages and

their linked pages, it was then realized that by taking the difference between the main page statistics and the average of the linked pages' statistics the author could form a two-space measure of how homogeneous the data was.

By doing so, it would be easy to say that if the page and its links scored a one that the content between the main page the linked pages was identical. Similarly if the main page metric was less than the linked pages' metric the score would be less than one and greater than one if the main page's metric scored better than the linked pages' metric.

By implementing that procedure the following statistics were derived from the Roleplayer data:

Table 5. Homogeneity Measurement of Roleplayer Population

	Fog Metric	Flesch Metric
Mean:	1.36177774434142	1.80675549532139
Standard Deviation:	0.743393027297314	1.54302620150437

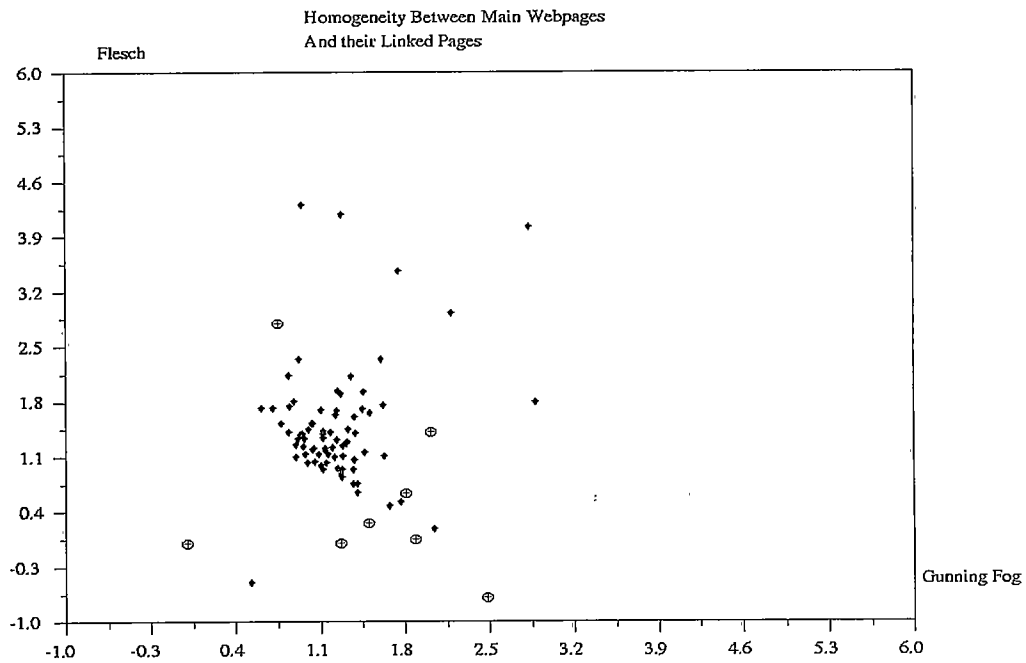
This data was, almost, what the author was looking for in terms of a low variability spread. It goes without saying that the data was starting to look like it wasn't going to be easily clusterable. But, as it wasn't 100% percent sure, a plot of the data was going to be a fair measure of what sort of clustering could be tried. In turn, this would indicate whether or not this would be a valid classification scheme.

FINDINGS

Phase I

Clustering in phase I isn't very good. This especially holds true for the Computer Science webpages, which are the circles with a cross in them. How ever note that Roleplayer community data shows a rather interesting trend not found in the Computer Science data. As seen by the measure of homogeneity, the roleplayers seemed to never link to a page that had a larger measure than their page. Another way to read this is that given a main webpage the linked webpages are, with the exception of a few outliers in the data, never better either in the Fog index or the Flesch index.

Figure 1. Homogeneity Between Webpages and their Linked Pages



Howsoever, as our intuition seems to corroborate, the Computer Science faculty tend to have very terse starting pages and in turn link to webpages that are instructional for their classes. In doing so they invariably show an opposite trend to that of the roleplayers, and have homogeneity scores less than 1 in either the Flesch score or the Gunning-Fog score.

Also of note is that the higher the main page scored the worse its links tended to be. This was due, in no small part, to gathering enough webpages that were incredibly high in score was error prone and did not yield enough samples to make a reliable estimate of how accurate this assumption was.

Phase II

Homogeneity in the secondary phase was more in line with the Roleplaying data of the primary phase of the research. We had gleaned 2396 samples in the initial phase and then carried out the calculation of the Average Information of their children and compared it.

As we had suspected prior, the homogeneity of the webpages for just the Information metric was generally greater than 1. As before, any measurement greater than 1 meant that the starting webpage had more information or scored higher than the children links in general.

As we can see in figure 2, in the smallest subset of pages found, the parents were still larger than the children were. In fact, further testing seems to indicate, via the graph, that as the majority of the webpages, 1997, were at least 1.1 times as great as their

children's average and that when this was extended out to a ratio of even 100:1 when we approached 1.

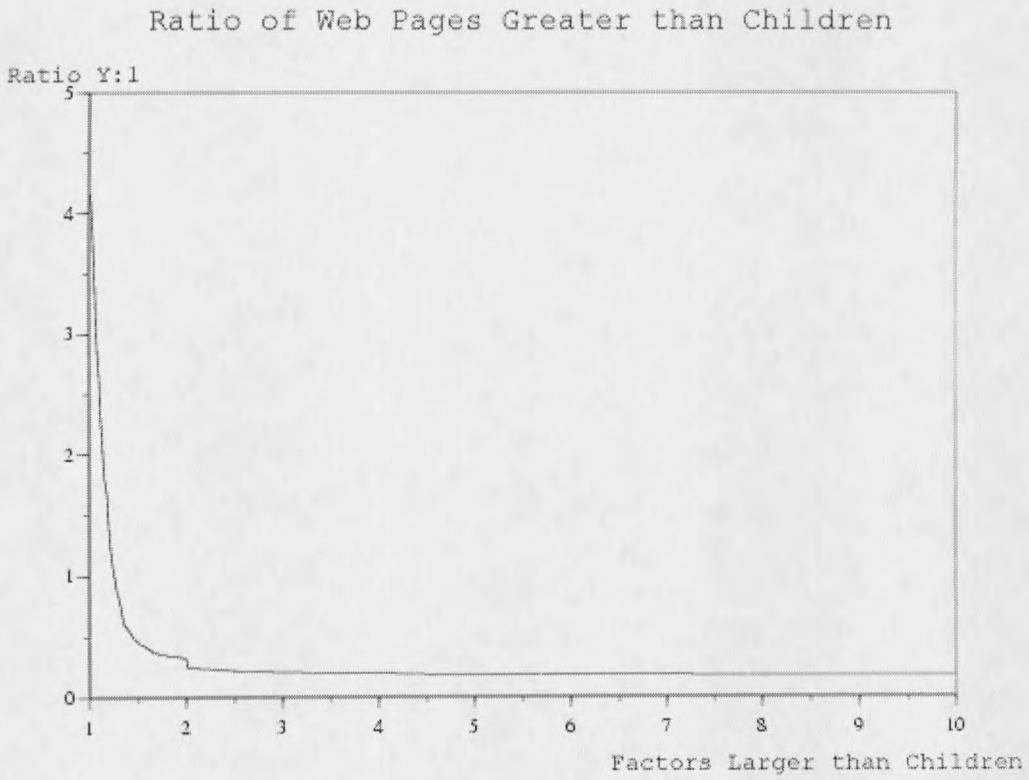


Figure 2. Homogeneity Ratio of Webpages Greater than Children

POST-ANALYSIS

Over the course of the study, it appears that people tend to not link to near-identical content as per our metrics. The original ideal of having a set metric X and then being able to make decisions on whether nor not to follow a link was rejected. Instead a measurement of homogeneity seemed to be more able to state some generalities of the webpages.

It also appears that there are not even some clusterable sets of data or classes that could be used to sufficiently explain or categorize the population in general. This holds to the fact that, mathematically, we could not pin down certain subsets of the population to classes such as "Flesch scores less than 50" and make general inferences about them.

In using homogeneity in metrics between parent pages and their children, we have seen that most sites tend to link to sites of a lesser content than their own page. If this holds true across generations, i.e. linked pages off of linked pages, we could form an algorithm that as a web engine's autonomous agents search for webpages for its database we could perform some expectancy function.

That function would be based on our data that points to a trend that if a webpage is of a sufficiently poor initial statistic set, such as scoring a 19 and 2, we could discontinue searching that page or its linked pages.

Similarly, due to our findings, we could design a function that if the metrics are particularly great, we'd discontinue searching down that path as well. This is due to the rapid decline in the score the linked webpages in our sample set tended to get.

In addition, we can see that academic pages tend in our very small sample set in phase I to reverse this trend so that the followed links are better than the original page and as such we'd search those out and categorize them first if at all possible. This is more than likely due to the metrics that were used, which were reliant on syllable counts and therefore could easily skew the score towards more complex word sets.

From all of this we can infer that if we were to start our own search engine we would rather start our search with pages that scored in the average of the metric set. This is counter to standard greedy search where we would expect that if we start with high scoring webpages we will always get high scoring linked webpages and therefore build a wonderful database. This is supported, in part, by figure 3.

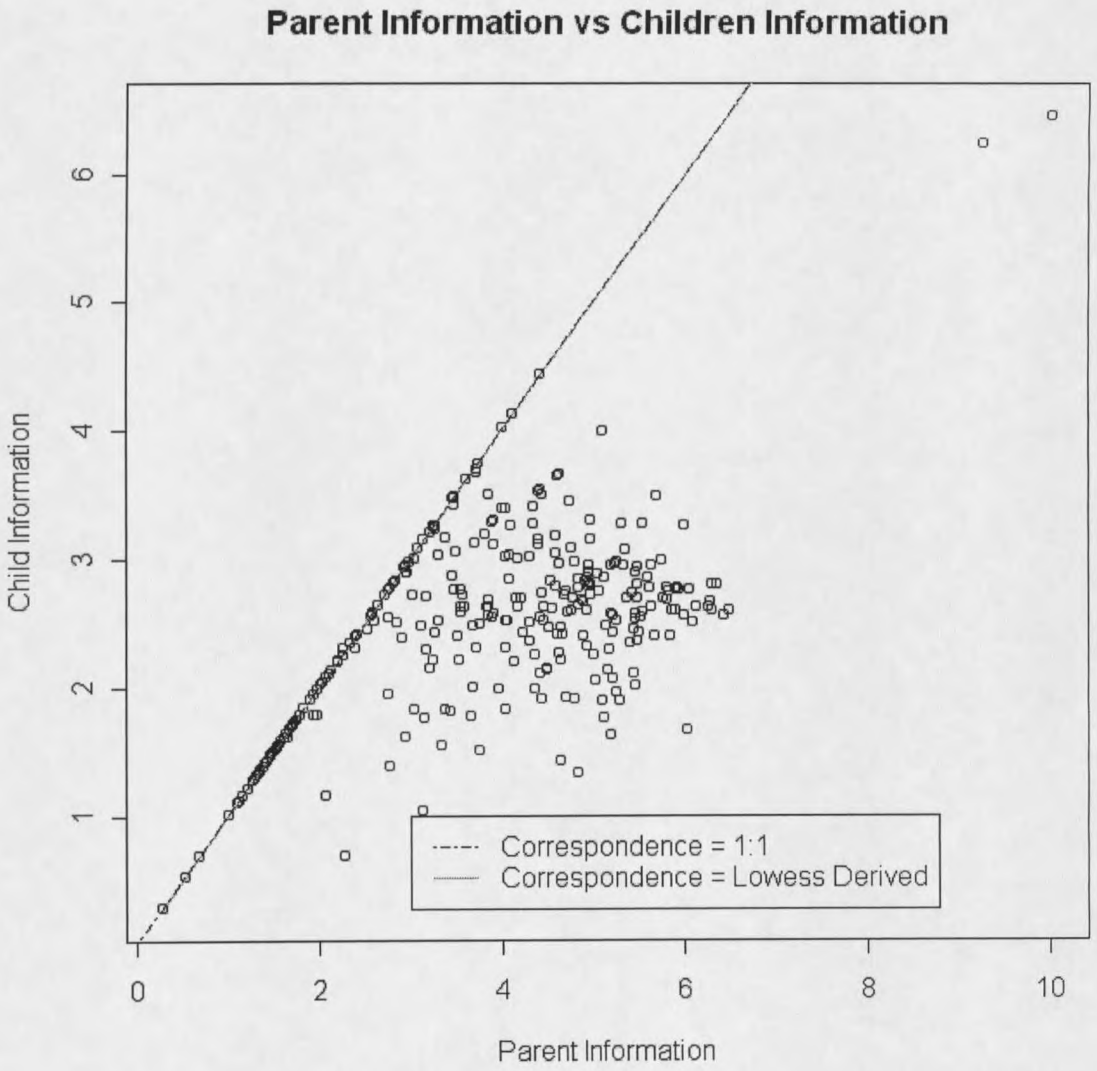


Figure 3. Parent Information versus Children Information

In figure 3, we see that, using information as an estimator, that given a least-squares estimate of the scatter plot (called a lowess derived line) that it appears that the majority of the webpages tend to lie on the line $y = x$. With this said, even though there the majority of the webpages tend towards having the same information measure as their children, it was a trend that if the parent scored over two that the children were either two or significantly less in score. Another interesting thing we can see is that given the graph

there were no children that scored a significantly higher score than the parent, which would have had a point in the area above the line $y = x$.

If we look at how this data shaped up, we see that due to the tree that is formed from just a few links and their linked articles that we need to be able to form some set of algorithms to pare the tree down. From what we've collected here we could significantly pare down the tree without losing much data just by performing some initial metric gathering on the initial limbs of the tree. In doing so, we would still retain significant data without much loss. Also, even if we score very high on one of those limbs we could pare it off after recording it down in our database.

For example, in our case, if three or higher was a significantly high scoring page, by applying some simple rule set on our starting tree we would have eliminated four pages. This doesn't sound like much but if our average links per page statistic holds we would have saved ourselves having to look at 140 new links. Since it took me 52 hours to statistically analyze 3518 links, we could analyze 1.13 links per minute. Therefore the savings would have roughly equated to 2 hours and 20 minutes of time shaved off our original search just by that rule.

CONCLUSIONS

In general our results confirm what we feel is common knowledge in that people don't tend to link to pages that outdo their own webpage. The only exceptions found where those of the Computer Science faculty webpages. These were people who were using their linked pages as instructional tools rather than other sources of comparable data and as such really couldn't be easily applied against the roleplayer group.

With this data in hand, there are a few conclusions that we can make. Since we have seen that people tend to cluster around like-metric pages, as a web search robot comes across webpages and given metric X, we could prune the search space.

This reduction of search space, would in turn allow for a deeper search in the tree that is the World Wide Web, and thereby potentially get more relevant topics to whatever we are hinging our search on. At worst, this allows for us to start a search on topics in a database that has a higher quality subset of pages.

In addition, with the results that people tend towards linking their webpages to other webpages of relatively similar value, we can include a lower bound so that if a webpage or set of webpages scores a particular value or lower then we don't follow their linked pages. This in turn also shortens the search space, and thereby allows us to retain some measurement of quality of information in the resulting search.

To recap, this paper has measured the homogeneity of parent webpages and the pages they link to. Through the examination of this information from the search on the

webpage a certain trend has been found that points to the idea that people don't seem to link to webpages that are more information- or other metric-dense than their own. As such, we can use these metrics as an expectation function to pare off parts of the search space, and as such reduce the number of webpages searched. This in turn could lead to a more pertinent or at least more information-dense webpage database.

REFERENCES CITED

1. Brin Sergey, Page, Lawrence. The Anatomy of a Large-Scale Hypertextual Web Search Engine. WWW7 / *Computer Networks* 30(1-7): 107-117, 1998

2. Flesch, Rudolph. *The Art of Readable Writing*, Harper and Brothers, 1949.

3. Gunning, Robert. *The Technique of Clear Writing*, McGraw-Hill, 1968.

4. Ryan, Kim, Lingua::EN::Fathom, http://www.cpan.org, 2000.

5. Salton, Gerard and McGill, M. *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983. ISBN 0-07-054484-0