Tracing Similarity within Strongly Connected Components for Intelligent Web Crawling

Yong-Jin Tee
Faculty of Computing and Informatics
Multimedia University Cyberjaya, Selangor, Malaysia
yjtee@mmu.edu.my

Lay-Ki Soon
Faculty of Computing and Informatics
Multimedia University Cyberjaya, Selangor, Malaysia
lksoon@mmu.edu.my

Abstract

Finding and obtaining information efficiently from the Web is one of the important elements in realizing Smart Home environment. Users expect to find most relevant information within the shortest possible time. In this paper, we investigate the similarity of Web pages within Strongly Connected Components (SCCs). SCCs are overlapping groups of Web pages that may imply a relationship between the Web pages of the same component. Therefore, we seek to trace the similarity of these groups of Web pages using Cosine Similarity. Our experiment performed on Malaysian Web pages indicates that Web pages within same SCC carry a common topic or theme. This finding proves that we may locate Web pages with similar topic using the hyperlinks structure, without performing expensive analysis on the contents of the Web pages.

1: Introduction

The Web can be viewed in the form of a graph, called the Web graph. The Web graph consists of nodes that represent Web pages, and directed links that represent the hyperlinks between them. Given the structure of the Web graph, we are able to extract groups of Web pages that are strongly connected to each other. These groups of Web pages are called Strongly Connected Components (SCCs). An SCC consists of Web pages where each Web page has at least a path to every other Web page within the SCC.

By extracting SCCs from the Web graph, we should be able to identify Web pages that maximally link to each other. Identification of such pages may facilitate Web page recommendation systems for Smart Homes, where Web pages that are related can be recommended to the user for further browsing. In this paper, we hypothesize that there must be a reason why the authors of these participating Web sites link their Web sites together. Therefore, SCC represents a relationship between the Web pages that could share a common topic. Directly crawling SCCs help to group together Web pages that are maximally connected.

The subsequent section describes some related work that uses SCCs in research. Later sections discuss the methodology that we apply to our dataset. After that, we report the preliminary results in our empirical study. We end this preliminary report noting our subsequent work in order to conclude our investigation.

2: Related Work

In recent years, literature involving Strongly-Connected Components (SCCs) pertain to the methods on extracting SCCs. These literature may be useful to us in the future, as the dataset may grow too huge for the existing tools used in our experiments to handle. In this paper, the size of the dataset is still managable by the tools for extracting SCCs. Therefore, we refrain from discussing literature on methods to find SCCs. Instead, we look to the others who use SCCs for various purposes.

Yamasaki [1] used SCC for topic extraction purposes. Yamasaki decomposed word co-occurrence graphs into SCCs to identify topics within documents. Saito et al. [2] studied link spam detection using SCCs. In their study of the Japanese Web, they found a considerable amount of link spams within their Web graph and discovered that most of their large SCCs contain link farms.

3: Methodology

We crawled a part of the Malaysian Web using Visioner-Bot[4], which is a highly customizable and flexible Web crawler. The dataset consists of a month's crawl of the Malaysian Web in May 2011. With the structured archives, we proceeded to construct the Web graph. The construction of the Web graph was done using Graphviz[1] . The SCCs were also extracted using the Unix package.

The data was cleaned using Jsoup[2] , which easily removed all hypertext tags. Stopwords were also removed from the text. Besides, to further reduce the complexity of the dataset, we reduced all words in every Web page into their stemmed forms using the Porter Stemmer [3].

For a SCC to be considered as a valid cluster of documents, participating Web pages within a single SCC should exhibit similar features. Therefore, these participating Web pages should bear a certain degree of similarity in terms of content. We used Cosine Similarity to gauge the similarity between pairs of Web pages in each SCC. The Cosine Similarity scores were then collected to calculate the Average Cosine Similarity Score (ACSS).

4: Results

4.1: The Dataset

There are a total of 165,962 Web pages and 236,690 hyperlinks in our dataset. These Web pages come from 6,470 different hosts. A total of 2,043 SCCs were found using GraphViz's sccmap. Among these 2,043 SCCs were 11,701 Web pages and 27,738 hyperlinks. We

[1] http://www.graphviz.org

[2] http://jsoup.org/

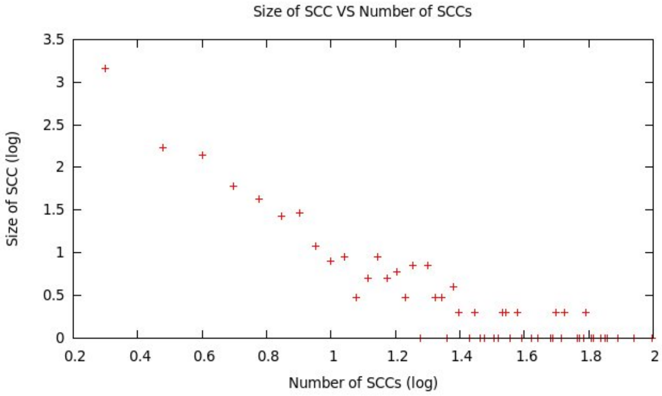

Figure 1. SCC Size Distribution for SCC Size less than 100

observed that the number of Web pages and hyperlinks in SCCs account for about 7% of total Web pages, and 11% of total hyperlinks. This goes to show that only a small number of Web pages will participate in SCCs. As such, visitors who stumbled upon any participating Web page may end up visiting Web pages within its SCC. These 7% contribute page hits to each other as Web surfers traverse through hyperlinks. Subsequent subsections will show example SCCs.

4.2: Distribution of SCC Sizes

We noticed that the SCCs vary in size. Intuitively we expect many SCCs that are small in size. After extracting all the SCCs, we inspected the size of all SCCs. We found that 1,442 SCCs have only two members. This makes for about 70% of all SCCs. Also, we found that the largest SCC has 1,756 members. After further inspection we discover that all of the 1,756 are Web pages from Olx[3] , a Web site that features classifieds advertising. The occurrence of large SCCs such as this is very rare. In fact there are only 8 out of 2043 SCCs having more than 100 members. For the sake of better readability, we show a graph where plots of SCC size more than 100 will not be displayed. Figure 1 shows the log-log plot for SCC size less than 100 versus the number of SCCs corresponding to the respective size. In this figure, the shape of the graph seems to obey the Power Law, where the graph displays a long tail. For our case, the graph suggests that large SCCs appear infrequently, while a majority of the dataset is made of small SCCs.

4.3: Distribution of SCC Sizes

We noticed that the SCCs vary in size. Intuitively we expect many SCCs that are small in size. After extracting all the SCCs, we inspected the size of all SCCs. We found that 1,442 SCCs have only two members. This makes for about 70% of all SCCs. Also, we found that the largest SCC has 1,756 members. After further inspection we discover that all of the 1,756 are Web pages from Olx[4] , a Web site that features classifieds advertising. The occurrence of large SCCs such as this is very rare. In fact there are only 8 out of 2043 SCCs

[3] http://www.olx.com.my/

[4] http://www.olx.com.my/

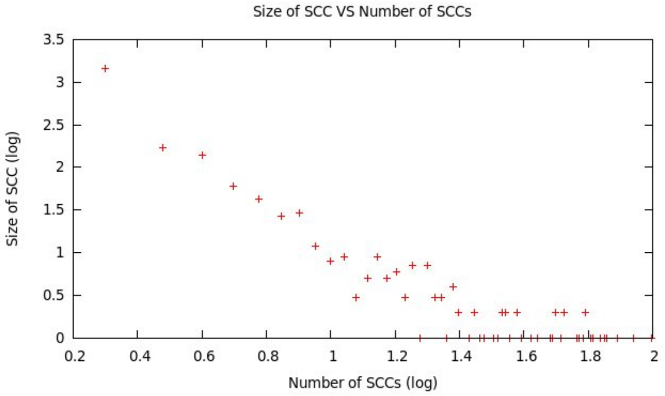

Figure 2. SCC Size Distribution for SCC Size less than 100

Table 1. Distribution of ACSS

ACSS	Number of SCCs	Percentage
0 - 0.1999	105	5%
0.2 - 0.3999	229	11%
0.4 - 0.5999	436	21%
0.6 - 0.7999	605	30%
0.8 - 1	668	33%
Total	2043	

having more than 100 members. For the sake of better readability, we show a graph where plots of SCC size more than 100 will not be displayed. Figure 1 shows the log-log plot for SCC size less than 100 versus the number of SCCs corresponding to the respective size. In this figure, the shape of the graph seems to obey the Power Law, where the graph displays a long tail. For our case, the graph suggests that large SCCs appear infrequently, while a majority of the dataset is made of small SCCs.

4.4: Average Cosine Similarity Score

To get an overall view of the distribution of ACSS, we first smooth all the ACSS scores to its bin boundary. We have chosen the bin boundaries in steps of 0.2 from 0 to 1. Table 1 shows the distribution of ACSS for all SCCs. It is shown that at least 95% of all SCCs scored a ACSS of at least 0.2. Therefore it can be said that most SCCs have a certain degree of similarity between Web pages that participate within the same SCCs.

Some SCCs have very high ACSS, especially if they originate from the same host. This is most probably due to the similarity in content and template design. An example of SCCs with high ACSS contains Web pages from the Mouth Cancer Awareness Week[5] Web site, which intuitively is about awareness of mouth cancer within Malaysia, and is hosted by University of Malaya. This SCC managed to obtain 0.71 ACSS. Figure 2 illustrates this

[5] http://mouthcancermalaysia.um.edu.my

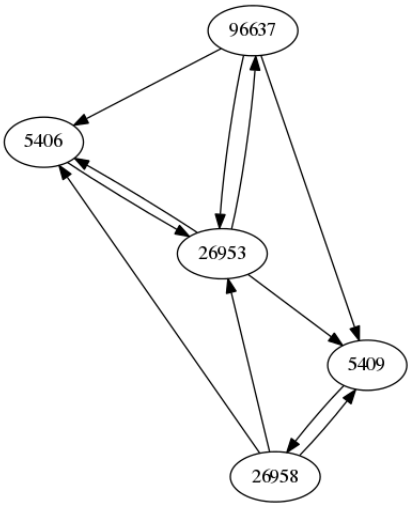

Figure 3. SCC 1850 that consists of Web pages from Mouth Cancer Awareness Week

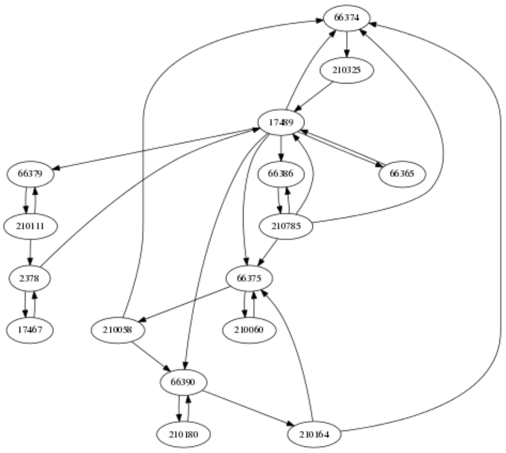

Figure 4. SCC 1604 that consists of Web pages relating to POS Equipments

SCC. The nodes are labelled with Web page IDs, where each ID corresponds to a Web page.

The largest SCC, which has 1,756 members, scored a low 0.28. Despite originating from the same Web site, we found that the pages contained varying content. This is due to the nature of the Web site, which features classifies advertising. Advertisers described their items using different terms and phrases that contributed to the low score. Large SCCs generally have lower ACSS than other smaller SCCs. Figure 3 shows an example of a moderately large SCC with a ACCS of 0.30. This particular SCC also contains Web pages from varying sites, which serves as another reason why the ACCS is low. Similar to the previous figure, the nodes are labelled with unique Web page IDs.

The SCC in Figure 3 is also interesting in the sense that even though the ACSS is low, all Web pages, except for 1, have a common topic. The odd Web page is actually the site of the Web designer for one of the other pages. This page linked back to the Web designer, thus the Web designer page is able to join the SCC. The common topic that the other pages have is POS (Point Of Sale) equipments.

5: Conclusion and Future Work

This paper reports on the preliminary study on Strongly-Connected Components (SCCs) as valid groups of Web pages containing a common topic or theme. We hypothesized that SCCs contain Web pages that are not only related on the structural level, but also on the content level. The average ACSS was found to be 0.65, and 1104 SCCs were found to be above average. The 1104 SCCs make 54% of total SCCs. This indicates that on average, SCCs lean slightly towards having high ACSS than having low ACSS.

In the near future, we plan to compare SCC with clustering results using primitive clustering methods. Given the information that the groups of Web pages may have a common topic, we have also planned on extracting associated terms[5] between Web pages participating within the same SCC. This branch studies on how these Web pages are associated, based on the prior hyperlink information that is being within the same SCC.

References

[1] Yamasaki, T.: Topic Extraction from Electronic Program Guides by using Decomposition of the Co-occurrence Graph into Strongly Connected Components In: Journal of the Database Society of Japan, Vol.7, No.1, June 2008, pp 1 - 6

[2] Saito, H., Masashi, T., Masaru, K., Kazuyuki, A.: A large-scale study of link spam detection by graph algorithms In: AIRWeb '07 Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web

[3] Jones, K. S., Willet, P.: Readings in Information Retrieval. San Francisco: Morgan Kaufmann, ISBN 1-55860-454-4.

[4] Qureshi, M., Younus, A., Rojas, F.: Analyzing the Web Crawler as a Feed Forward Engine for an Efficient Solution to the Search Problem in the Minimum Amount of Time through a Distributed Framework. In: International Conference on Information Science and Applications (ICISA), pp 1 - 8

[5] Tee, Y. J., Soon, L. K., Ranaivo-Malançon, B.: Finding Web Document Associations Using Frequent Pairs of Adjacent Words In: Proceedings of 3rd Semantic Technology and Knowledge Engineering Conference (STAKE), pp 57 - 60