Structure-Based Analysis of Web Sites

Benjamin Yen
School of Business, The University of Hong Kong, Pokfulam Road, Hong Kong
benyen@business.hku.hk

Abstract

The performance of information retrieval on the Web is heavily influenced by the organization of Web pages, user navigation patterns, and guidance-related functions. Having observed the lack of measures to reflect this factor, this paper focuses on an approach based on both structure properties and navigation data to analyze and improve the performance of Web site. Two types of indices are defined two major factors for analysis and improvement – "accessibility" reflects the structure property to measure how easy the user can access the pages and "popularity" implies the navigation data primarily based on the log statistics. The accessibility and popularity (A-P) plot serves as a compass for the Web designer to get an overview of current performance status and explore in the possible directions for improvement to balance the design anticipation and navigation expectation.

1. Introduction

The Internet has changed the way people communicate with each other as well as expedited the process to obtain the information that matches their interests. Together with the growth of the needs of information, the web pages on the Internet grow explosively during the past few years and such increase is expected to be more acute over time. However, too much unorganized information may create problems with searching performance in the meanwhile. Visitors spend astounding amount of time in navigating through the useless or redundant pages. As a result, visitors and web page owners are facing the same problems: How to balance the gain and the cost so that both can get the most satisfaction? How can the Web site serve the visitors better for information retrieve?

Many researchers have worked on Web information retrieval in the recent years. Perkowitz and Etzioni [1] describe two cluster-mining algorithms that gather Web pages that are not linked but related to the same topics in the user's mind automatically. Nakayama *et al.* [2] address a technique that determines the gap between Web site designers' expectations and users' behavior. Joachims *et al.* [3] introduce a learning approach with the user feedback to improve the quality of advice for navigation interactively. Chakrabarti et al. [4] develop algorithms that exploit the hyperlink structure of the WWW for information discovery and categorization. Sarukkai [5] uses a Markov chain model based on the user access information for link prediction and path analysis. Gibson et al. [6] define the Web sites as "authorities" and "hub" in isolation and conclude that a respected authority is page that is referred to by many good hubs and a useful hub is a location that points to many valuable authorities. Zin and Levene [7] propose that information on the topology is important for useful exploration. Gilleson et al. [8] establishes taxonomy of Web site traversal patters and structures. Dhyani et al. [9] classifying and discussing a wide ranging set of Web metrics for quantifying various properties to improve Web information access and use. Garofalakis and Mourloukos [10] define a relative popularity based on the absolute popularity, page depth, number of pages on the same level and number of incoming hyperlinks.

2. Modeling

The accessibility of an object (a page, for instance) is determined by a variety of factors, such as its location in a Web site and the links pointing to it. The efficiency of information retrieval is closely related to the Web structure. Analysis of the accessibility of Web pages can help the Web designer improve the quality of link structures. The accessibility of Web links and pages is defined to measure the availability of a link in a page and a page on a site respectively. The relationship, namely *relevancy*, between the popularity and accessibility are analyzed to discover the potential problems for further improvement on Web structure.

0-7695-2073-1/04 $20.00 © 2004 IEEE

The measurement of how easy the user can access the pages can be denoted as average time to access the page, the percentage of time user will stay at the page, or estimation of easiness to access the page. Four exemplary accessibility models based on the properties of graph structures are proposed, namely *Expected Link Number Model*, *Accumulated Accessibility Model*, *Sum of Distance Reciprocal Model*, and *Sum of Expected Distance Reciprocal Model*. These models are based on the accessibility of links and pages, and other properties, such as expected value, accumulation and distance [11].

The alternative of the definition for accessibility can be the reciprocal of average access time from the entrant page to the destination page. To calculate the average navigation time, we need to consider both the access time of the target page and the access time (i.e. opportunity cost) of the other pages probably visited before the target page. The average access time is sum of access time of the target page and expected access time of other sibling pages as

$$E[N_t] = T_t + \sum(T_a + \tfrac{1}{2}\sum S_i) = T_t + \sum(T_a + \tfrac{1}{2}\sum\sum T_{ij}).$$

where $E[N_t]$ is the expected (average) access time of the target page t, T_t is the access time of the page t, T_a is the access time of the immediate ascendant node, S_i is the sum of the access time of all nodes descendent to node i (the sibling of target node), and T_{ij} is the access time of the j^{th} descendent of sibling page i. If the target page is on the path of a cycle, we can exclude the immediate ascending pages whose level is smaller than that of the target page since we only consider the average access time of the first visit to the target page.

The importance and accessibility of a page should be co-related. The "popularity" of a Web page can be regarded as the importance of that page, which can be measured in terms of "views", the number of times a page is viewed by the users, or "visits", the number of times a page is retrieved in different sessions. Other measures include average time the users spent on a page, percentage of total views/visits, etc. In the case study, the popularity is measured in "visits". However, the user preference may vary over time and beyond the Web designer's expectation. The Web designer, thus, should continuously examine the users' navigation behavior and update the association of the Web pages.

The *A-P* (*A*ccessibility-*P*opularity) plot can help discover the gap between the Web designer's anticipation and the users' expectation. Fig. 1 shows the relevancy relationship in *A-P* plot. There are four zones I, II, III, and IV in the plot to represent different types of the pages. The zone I includes the pages with high accessibility and high popularity which are the most preferable cases for *A-P* consistency in a Web site. The pages in zone II have high accessibility and low popularity where the pages may need to be pushed to the "corner" in order not to take up the too much of the "prime" area. The zone III consists of pages with low accessibility and low popularity where these pages are either newly launched for adjustment or obsolete and ready for removal. The pages in the zone IV have low accessibility and high popularity and their accessibilities may worth being increased.

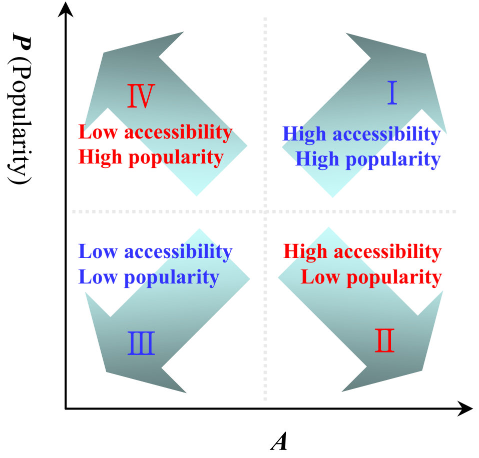
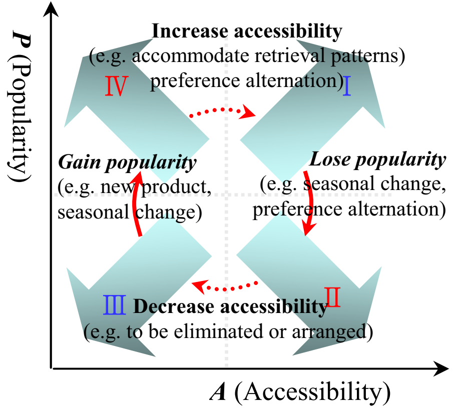


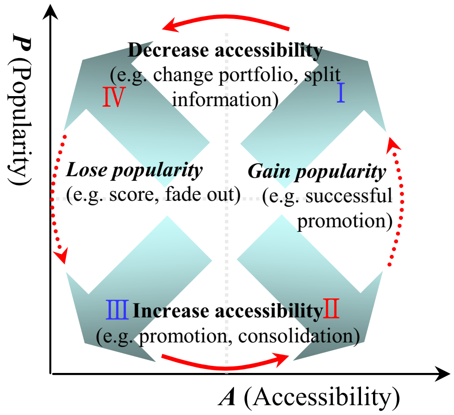
Fig. 1 *A-P* (Accessibility-Popularity)

In the *A-P* plot, the *x*-axis (accessibility) represents the direction that Web designers can maneuver and the *y*-axis (popularity) depends on the navigation patterns of the user. The evolution of pages on A-P plot can be in a Push cycle or a Pull cycle. In a *Push cycle* as shown in Fig. 2, the evolution is based on the user demand. The decision on the changes in accessibility is dependent on the popularity reflected in user preferences. The pages in zone I (high accessibility and high popularity) might move into zone II due to the changes in the user preference and the designer may decrease their accessibility to move them into zone III for further action later. If they stay in the zone III, they can be removed as obsolete pages; or they become popular again to move into zone IV, then the designer can increase their accessibilities to move them into zone I. In a *Pull cycle* as shown in Fig. 3, the evolution is based on the site design. The increment in accessibility is expected to lead the improvement on the popularity. The pages in zone III can be promoted or consolidated by increasing their accessibility to move them into zone II, and they are expected to gain attention and popularity from the users to move into zone I. Sometimes, the site designers need to change the page portfolio (such as to keep the ten most important pages) or split some information to adjust the site balance. In this case, the pages in zone I move to zone IV. Some pages in zone IV may gradually lose the popularity and move to zone III to be eliminated. Both Push cycle and Push cycle can be used as guideline for Web site management.



P (Popularity)
Increase accessibility
(e.g. accommodate retrieval patterns)
preference alternation)
IV
I
Gain popularity
(e.g. new product, seasonal change)
Lose popularity
(e.g. seasonal change, preference alternation)
III Decrease accessibility II
(e.g. to be eliminated or arranged)
A (Accessibility)

Fig. 2 *A-P* Evolution – Push cycle



P (Popularity)
Decrease accessibility
(e.g. change portfolio, split information)
IV
I
Lose popularity
(e.g. score, fade out)
Gain popularity
(e.g. successful promotion)
III Increase accessibility II
(e.g. promotion, consolidation)
A (Accessibility)

Fig. 3 *A-P* Evolution – Pull cycle

3. A-P Analysis

Generally there are two ways to improve the accessibility of a Web page (or a group of pages): (1) increase the level of the page (group), or (2) add more links pointing to the page (group) in other pages. The pages with high popularity have the higher priority for accessibility improvement. This is achieved by checking the relevancy between the accessibility and the popularity. The procedure needs to be repeated continuously since the popularity or "importance" of the Web pages changes over time. The accessibilities can be determined from the Web site itself while the popularities can be obtained from the log file. Based on the relevancy, problems may be found and the accessibilities of the Web pages can be improved accordingly. The procedure can be repeated routinely. A case study based on Expected Link Number (EN) model can be found in [11].

In order to decide the accessibility for each page, we re-examine A-P plot shown in Fig. 4. The diagonal line going through zone I and III represents the *balance* line for accessibility and popularity. This line also serves as the expected baseline to decide or adjust the accessibility of Web pages. For each page, we can select a targeted reference point on this line. For example, point *A* in Fig. 4 represents a page with very high accessibility and popularity which is suitable for the most important or highly-promoted pages. On the other hand, point *D* denotes the expected position for some minor pages. The horizontal direction represents the adjustment of Web structure and the vertical direction reflects the changes of navigation results. The procedure to decide the page accessibility is as following:

(i) Choose a targeted point on the balance line and we can decide the initial accessibility.

(ii) After the testing period, we can collection information for the popularity for this page to get the actual A-P value on the A-P plot and we can find a new reference balance point to get a updated reference accessibility.

(iii) We can choose the value between the original and updated accessibilities to be the new accessibility.

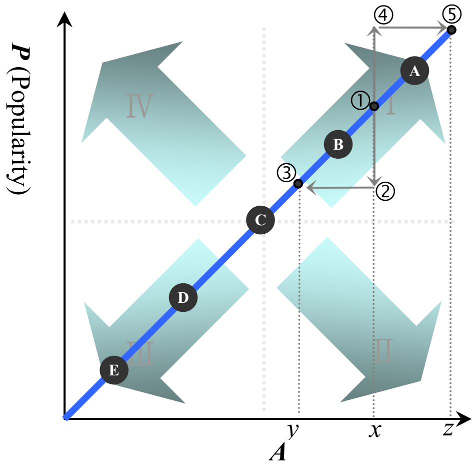


Fig. 4 *A-P* segmentation

Consider the scenario that we add a new page as shown in Fig. 4. Assume the initial balance point is at point *1* and the corresponding value of accessibility is projected at x-axis, *x*. After launching the page for a period of time, if the actual A-P is at point *2*, then the new possible balance point becomes point *3* and the new corresponding accessibility is *y*. We may choose the any value between *x* and *y* as the new accessibility for the page. If the actual A-P is at point *4*, we can find point *5* as a new reference point and the corresponding accessibility is *z*. Similarly, we can choose any value between *x* and *z* as the new accessibility. We can continue the process in a similar procedure.

The A-P analysis involves two important issues – "What pages should be considered for A-P analysis?", "How to improve A-P definition?", and "How to use the A-P analysis for to improve the performance?"

A. What pages should be considered for A-P analysis?

There are three page groups critical for the A-P analysis -

(i) *Leaf nodes*. The main information (such as product information) is located in the leaf nodes.
(ii) *All nodes*. All the nodes will be included for analysis.
(iii) *Selected nodes*. The selection can be path-based (e.g. navigation pathway), level-based (e.g. electronic catalogs), function-based (e.g. taxonomy), etc.

B. How to use the A-P analysis to improve the performance?

The organization of Web pages can be modeled as a *conceptual network*. The structure of the Web site is an instance of the conceptual network. A conceptual network consists of the information entity, information composition, and information constraints. *Information entity* is the content of the raw information. *Information composition* describes intra-entity relationship, such as the format and arrangement of the information; *information constraints*, on the other hand, define inter-entity relationship, such as precedence, serial/parallel linkages. The constraints can be further classified as soft constraints and hard constraints.

The improvement may involve both critical and non-critical pages in A-P analysis. The result of A-P analysis should be better than the original pages setting. Side-effect, such as the tremendous changes in the non-critical pages or violation of information constraints, should be avoided. Some guidelines of improvement process are as follows -

(i) Start with the most critical pages.
(ii) Swap the pages in the opposite directions in the A-P Plot.
(iii) Follow the bottom-up (deep leaves) approach.

However, the guidelines may not guarantee the improvement expected in the general structure due to the cascading effect, oscillation effect, etc.

C. How to improve A-P definition?

The classification of four zones in A-P plot lies at the threshold values of accessibility and popularity on x-axis and y-axis respectively. The threshold values can be decided by statistical method (e.g. average), heuristics (e.g. neural network), or benchmark models.

4. Conclusions

In order to design and maintain a friendly Web site, the Web designers need to decide how to adjust the Web structure in a dynamic access environment. The *A-P* plot shows the classification of the pages in a Web site to guide the site designers to manage the site in either a Push cycle or a Pull cycle. The following issues need to be resolved as further research directions:

(1) *Improvement guideline*. Needs more investigation in the improvement guidelines to avoid the anomalies and oscillation.
(2) *Time-based analysis model*. Include the page size or navigation time for each page.
(3) *Group-based analysis model*. Include the page group of interest or pages on the same path.

5. References

[1] M. Perkowitz and O. Etzioni, Towards adaptive Web sites: Conceptual framework and case study. *Artificial Intelligence*, 118, pp. 245-275, (2000).

[2] T. Nakayama, H. Kato and Y. Yamane, Discovering the Gap Between Web Site Designers' Expectations and Users' Behavior. *Proceedings of WWW9*, Amsterdam, Netherlands, (2000).

[3] T. Joachims, D. Freitag and T. Mitchell, WebWatcher: A Tour Guide for the World Wide Web. *Proceedings of IJCAI-97*, Nagoya, Janpan, pp770-775, (1997).

[4] S. Chakrabarti, B. Dom, S.R., Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson and J.M. Kleinberg, Mining the Web's Link Structure. *IEEE Computer*, 32(8): 60-67, (1999).

[5] R.R. Sarukkai, Link Prediction and Path Analysis using Markov Chains. *Computer Network*, vol. 33 pp377-386, (2000).

[6] D. Gibson, J. Kleinberg and P. Raghavan, Structural Analysis of the World Wide Web. *WWW Consortium Web Characterization Workshop*, November, (1998).

[7] N. Zin and M. Levene, Constructing web views from automated navigation sessions. *ACM Digital Library Workshop on Organizing Web Space (WOWS)*, Berkeley, 54-58, (1999).

[8] M.L. Gillenson, D.L. Sherrell and L. Chen, A Taxonomy of Web Site Traversal Patters and Structures. *Communication of the Association for Information Systems*, Vol. 13, Article 17, (2000).

[9] D. Dhyani, W.K. Ng and S.S. Bhowmick, A Survey of Web Metrics. *ACM Computing Surveys*, Vol. 34, No. 4, pp. 469-503, December, (2002).

[10] J. Garofalakis and D. Mourloukos, Web Site optimization Using Page Popularity. *IEEE Internet Computing*, July-August, pp. 22-29, (1999).

[11] B.P.-C. Yen, The Design and Evaluation of Accessibility on Web Navigation. *Proceedings of ICEC*, Hong Kong, (2002).