Inferring Sub-culture Hierarchies Based on Object Diffusion on the World Wide Web

Ta-gang Chiou, Judith Donath
Media Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02142, USA
Tel: 1-617-253-0323
E-mail: {dc,judith}@media.mit.edu

Abstract

This paper presents our approach to inferring communities on the Web. It delineates the sub-culture hierarchies based on how individuals get involved in the dispersion of online objects. To this end, a relatedness function is chosen to satisfy a set of carefully defined mathematical conditions. The conditions are deduced from how people may share common interests through placing common objects on their homepages. Our relatedness function can infer much more detailed degree of relatedness between homepages than other contemporary methods.

1. Introduction

1.1. Cultural dispersion on the Web

On the World Wide Web (WWW), people embellish their homepages with links, pictures, sounds, etc. The spread of these virtual objects is part of the cultural dispersion on the Web. In a social hierarchy, the dispersion is possibly fueled by the forces of differentiation and emulation, according to the sociological “trickle-down theory [1],” and there can be many such hierarchies. The dispersion may reveal a lot about the social structure of a society - it shows where there may be contact between individuals and delineates the sub-cultural hierarchies. These sub-cultures are developed by people who have similar runs of experience through common interactions, according to the sociology of “collective selection [2].”

We are doing research on tracking and analyzing cultural dispersion on the Web[1], which deals with many of these issues. One of the big pieces needed for the research is to be able to measure how homepages are related to one another based on the objects they have adopted from online cultural dispersion. For this purpose, we develop a relatedness function, as will be discussed later, which infers the relatedness between homepages from the number of common objects, including links, images, and so on.

1.2. Inferring communities from link information

As the WWW grows in size and complexity, inferring high-level structure on the Web becomes increasingly important. There has been a growing amount of work [3,4] directed at the integration of textual content and link information to infer structures of communities on the Web.

Pitkow and Pirolli [5] found that co-citation analysis [6] can be helpful for identifying interesting clusters of pages on the web. The co-citation analysis has been very helpful for categorizing scientific papers according to how articles cite one another. In fact, both studies in co-citation and bibliographic coupling [7] can be inspiring to infer relatedness between pages. For two documents p and q, the co-citation quantity is equal to the number of documents cited by both p and q, and the bibliographic coupling quantity is the number of documents that cite both p and q. The larger these quantities are, the more likely p and q are about research in the same field.

Gibson and Kleinberg [8] try to find interesting communities of pages on the Web through an analysis of link topology. The communities can be viewed as containing a core of central, "authoritative pages" linked together by "hub pages"; and they exhibit a natural type of hierarchical topic generalization that can be inferred directly from the pattern of linkage.

Terveen and Hill [9] define a structure, called the clan graph, that groups together sets of closely connected sites

[1] The project homepage is at http://vfashion.media.mit.edu

from a set of seed documents based on connection between them. The clan graphs treat links as undirected.

1.3. Our approach - developing a relatedness function based on cultural dispersion

In this paper, we present our method for inferring communities on the Web. It delineates the sub-culture hierarchies based on how individuals get involved in the dispersion of online objects. To this end, a relatedness function is derived from a set of carefully defined mathematical conditions. The relatedness function can infer much more detailed relatedness between homepages than previous methods, which use simple relation, either binary (connected or unconnected) [9] or use the sum of common links [5]. It helps to generate meaningful group hierarchy even if the node-to-node distance of pages is not taken into account.

Moreover, we differentiate homepages from objects on them, instead of mixing them all in a hypertext structure as in [8] and [9], so that we can take into account virtual objects, whether they be HTML documents or not.

Finally, the relatedness function can work well independently of whether a page or a site is chosen to be the basic unit, so that the program may discover communities of documents, as inferred by [5,8], as well as communities of people. In the context of the larger area this project addresses, the communities of people mean a lot more than communities of documents in that they may indicate further areas of research about particular sociological aspects of the Web.

2. Inferring communities of cultural dispersion

2.1. The basic hypothesis

Inspired by research on co-citation analysis [6], we can begin to deduce the relatedness of homepages from the number of common objects, such as links to the same set of URLs. In this paper, we will use hypertext links as the objects we refer to for the purpose of simplification.

The intuition is that if two homepages both link to the same set of URLs, their owners may share similar interests and thus may be involved in the same sub-culture. The more links they have in common, the more likely the owners share similar interests.

In other words, we start with counting the co-citation quality of two homepages, but not the bibliographic coupling quantity [7]. This is because the bibliographic coupling quantity is difficult to calculate precisely unless we have a complete set of the WWW, and it cannot show the shared interests of the authors.

For simplicity, in this paper, we will regard a homepage as the online identity of its owner. Thus, “homepage A and B may share the same interests” means that their owners may share the same interests. Note that in this paper we specially regard a homepage as the site a certain individual owns, instead of another meaning, such as "the first page," of the term "homepage." Therefore, the term "homepage" in this paper means the same as the new term "homesite" used by some web developers.

2.2. Mathematical conditions

First, we need a function for quantitatively inferring the relatedness between any two homepages. This should not be a binary (connected/unconnected) relation, because a binary value discards too much useful information about different degrees of relatedness.

The simplest function one may come up with is to calculate the total number of common links between two homepages, as used in traditional co-citation analysis [6] and [5]. Although it works well for bibliography, this method is inaccurate for analyzing the Web since homepages vary significantly in size. This is resulted from the fact that publishing hundreds of papers is quite difficult, but putting hundreds of links on a homepage is relatively easy. Thus, by counting only the total number of common links, portal sites may be highly related to most homepages. It is because portal sites contain so many links that they tend to have links in common with any other sites, while in fact portals are not especially affiliated with most of these homepages.

On the basis of this observation, one may refine the function to divide the number of common links by the number of total links on each homepage. Our experiments show that this method results in very distorted and unfavorable measurements: the denominator (the number of total links) has such a great effect that the numerator (the number of common links) rarely makes a difference. The size of the homepage turns out to be the dominant, and almost the only, factor.

To derive a carefully designed function, a list of mathematical conditions should be helpful. To begin with, let us think of the affiliation of a homepage A to any URL it links to. If A links to t_A URLs in total, it is intuitive that the smaller t_A is, the higher the affiliation A may have to any URL it links to. For example, if both homepages A and B link to http://www.media.mit.edu but A links to a total of 10 URLs and B links to a total of 1000 URLs, it is likely that A is more affiliated with http://www.media.mit.edu than B.

Based on the idea of affiliation, we can think of the relatedness between homepage A and homepage B. Again, assume that A links to t_A URLs in total, and B links to t_B URLs in total, and they have c_{AB} links in common (an example is given in Figure 1.) It is obvious that A and B are more likely to share the same interests when c_{AB} becomes larger or when t_A and t_B get smaller.

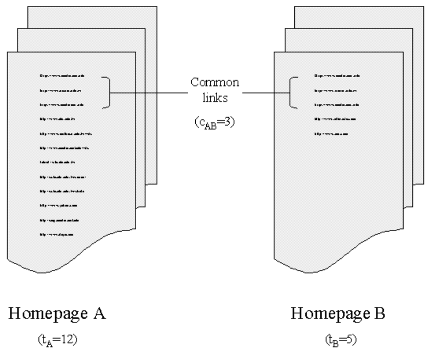

Figure 1. An example of having common links between homepage A and homepage B.

To simplify the problem, we may list the mathematical conditions for calculating the given homepage A's affiliation to the number of common links, c_{AB}. Later, by combining both A's and B's affiliation to the set of common links, we can get the final relatedness function between A and B.

As listed below, we have integrated the mathematical conditions for selecting the affiliation function which calculates A's affiliation to the links A and B have in common. Let's call the affiliation function $f(c_{AB}, t_A)$. The first parameter is the number of the links, which are linked by both A and B, we want to count homepage A's affiliation to. The second parameter is the total number of links on homepage A. The parameters are always natural numbers.

(1) $f(n, m) < f(n+1,m)$, where $n+1 < m$.

The larger the number of common links, the greater A's affiliation to this set of links.

For example, two out of three links on A implies more affiliation than one out of three links on A.

(2) $f(n, m+1) < f(n, m)$, where $n < m$.

The larger the number of total links on A, the less A's affiliation to a certain set of links on it.

For example, one out of three links on A implies more affiliation than one out of one hundred links on A.

(3) $f(n, m) < \mathrm{f}(an, am)$, where $n < m$, and $a > 1$.

Even though the fraction n/m is equal to an/am, the larger the numbers, the more concrete evidence of affiliation.

For example, three out of nine links on A implies more affiliation than one out of three links on A.

(4) $f(n, m) < a\,f(n, am)$, where $n < am$, and $a > 1$.

From our experience, the decline of the value of affiliation should not be too reactive to the total number of links, as the lower curve on Figure 2. Otherwise, the total number of links will be the only dominant factor of the function. A less steep curve, the upper curve on Figure 2, may resolve this problem. Thus, $f(n, am)$ should be larger than $(1/a)$ times $f(n, m)$.

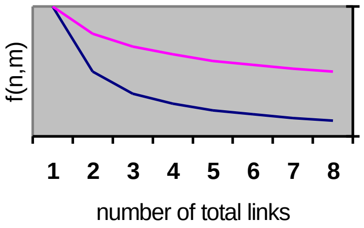

Figure 2. Possible change of f(n,m) when n=1. The lower line represents a function that reacts too much to m, the number of links on the homepage in total.

2.3. Relatedness function

There can be many possible functions to satisfy these mathematical conditions. The affiliation function we choose is:

If n=0,

$f(n, m)$=0

else

$$f(n, m)= \frac{n^{\frac{1}{r_n}} + k_n}{m^{\frac{1}{r_m}} + k_m} \quad \text{,where } 0 < k_n \le k_m,\ 1 \le r_n < r_m.$$

r_n is the order of the root taken of n, r_m is the order of root taken of m.

The reason that r_m should be larger than r_n is that m is always larger and varies a lot more than n. To avoid m becoming the dominant factor, r_m needs to be larger than r_n. The coefficients: k_n, k_m, r_n, r_m can be determined by the feedback from the actual data. Our experiment shows that to set r_n around 1 and $r_m \ge 2$ works well.

Examples of refining the coefficients are shown in Figure 3 and Figure 4.

Thus, the relatedness function between A and B can be labeled $g(f_A, f_B)$. We can then set, for example,

$$g(f_A, f_B) = \sqrt{f_A f_B}\,,$$

which is the geometric mean of A's and B's affiliation to their common links. This function is one of the simplest functions that fit the mathematical conditions we specified. As long as a function fits the conditions and has appropriate coefficients, it should behave reasonably well.

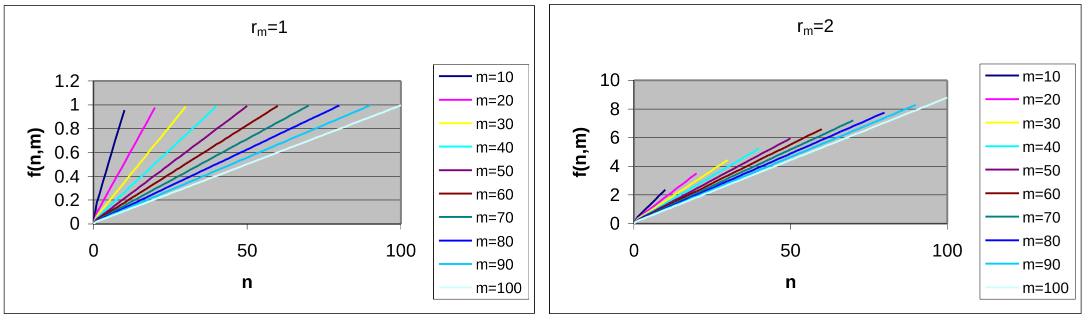

Figure 3. The left and right charts show samples of the behavior of the affiliation function for r_m=1 and r_m=2, respectively. The right chart shows a more favorable choice of rm.

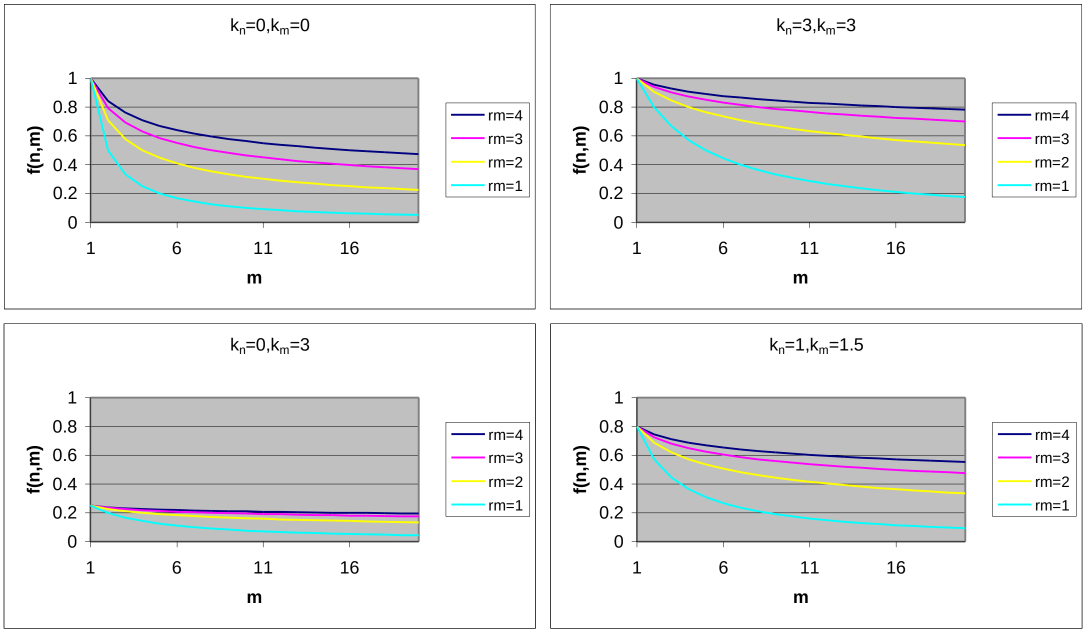

Figure 4. For sample case of n=1, above charts show the behavior of the affiliation function for different k_n and k_m. The bottom-right one is more favorable than others.

2.4. Discrimination value

For better accuracy, we also refine the function in our program based on some other criteria. For instance, instead of treating all links on homepages equally, weighting links that are shared by only a few people more can be more accurate. Links that are linked by only a few people may reveal the specific interests of their users better than those widespread links. For example, almost all homepage in the GeoCities have a link to the GeoCities front page, and only a few homepages have a link to a certain group in the MIT Media Lab. Obviously, the latter reveals a lot more about a person's specific interest than the former.

Besides, incorporating time as a factor may make the underlying information flow even clearer. Since a person may change his interest, the link a person recently links to can reveal more about his recent interest. The length of the period over which an individual adopts a certain link may also be meaningful: short-period means short-term interest. Furthermore, some links may get popular in a short period of time. These recently fashionable links may have significant cultural meanings so that they get popular so fast. Therefore, the fashionable links can be weighted more than ordinary links. The statistics of the temporal dynamics can be obtained from a system called "virtual fashion" that we have developed.

Based on these observations, the relatedness function can be further refined to give every link different discrimination value. The actual implementation differs according to which criteria are more important for the specific research.

2.5. Clustering the homepages based on their relatedness

By calculating the multidimensional relation between homepages with the relatedness function, our program categorizes homepages into hierarchical clusters [10]. The relatedness is equal to the "distance" from the perspective of clustering algorithms. Starting from the highest relatedness observed, the program progressively decreases the threshold of relatedness to find homepages connected directly or indirectly to one another with relatedness above the threshold. Thus it gradually discovers sub-culture hierarchies within the set of homepages. Efficiency can be increased by optimizing the implementation of the clustering algorithm, a discussion of which is beyond the scope of this paper.

Our approach generates meaningful group hierarchies even if the node-to-node distance of pages is not taken into account. Our experiments with about nine thousand homepages show satisfactory results. For example, in the MIT Media Lab domain, a group inferred by the method usually consists of people in similar research fields, and a certain sub-group may consist of people in the same research team.

3. Conclusion

By analyzing the dispersion of online objects among homepages using a relatedness function, we develop a method to infer sub-culture hierarchies on the Web. The relatedness function may also be incorporated into other contemporary approaches such as [5,9] to refine their measure of similarity/relatedness between pages.

In addition to clustering homepages, the function can also be applied to classifying virtual objects. The hypothesis then becomes: if two objects are both linked by the same set of homepages, they may be related. By tracking and analyzing cultural dispersion on the Web, we hope to further understand how the Web functions as a social environment.

4. Acknowledgements

This research is part of the project "Virtual Fashion - Tracking and Analyzing Cultural Dispersion on the World Wide Web."

We thank Kan Liu, Wen-Jung Tsai, Roy Rodenstein, Benjamin Vigoda, Ching-Huang Yun, Ben Fry, Ta-Lun Chiou, Andreas Weigend, and Michael Best for their valuable input to this research.

5. References

[1] Simmel, Georg. "Fashion." Rpt. In American Journal of Sociology 62 (May 1957): 541-558. 1904.

[2] Blumer, Herbert. 1969. "Fashion: From Class Differentiation to Collective Selection," Sociological Quarterly, 10(3), 275-291.

[3] R. Botafogo, E. Rivlin, B. Shneiderman, "Structural analysis of hypertext: Identifying hierarchies and useful metrics." ACM Trans. Inf. Sys., 10 (1992). 142-180.

[4] P. Pirolli, J. Pitkow, R. Rao, "Silk from a sow's ear: Extracting usable structures from the Web." Proc. ACM SIGCHI Conference on Human Factors in Computing, 1996.

[5] P. Pirolli and J. Pitkow. "Life, Death, and Lawfulness on the Electronic Frontier." Proc. 1997 Conference on Human Factors in Computing Systems, CHI 97, ACM, Los Angeles, CA, USA. 383-390.

[6] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents." J. American Soc. Info. Sci., 24(1973), 265-269.

[7] M.M. Kessler, "Bibliographic coupling between scientific papers," American Documentation, 14(1963), 10-25.

[8] D. Gibson, J. Kleinberg, P. Raghavan. Inferring Web communities from link topology. Proc. 9th ACM Conference on Hypertext and Hypermedia, 1998.

[9] Terveen, L. and Hill, W, "Finding and Visualizing Inter-site Clan Graphs." Proc. 1998 Conference on Human Factors in Computing Systems, CHI 98. ACM, New York, NY, USA. 448-455.

[10] Everitt, B., "Cluster Analysis," Halsted, NY, 1980.