Subgroup Discovery among Personal Homepages

Toyohisa Nakada and Susumu Kunifuji

Japan Advanced Institute of Science and Technology,
Tatsunokuchi, Ishikawa 923-1292, Japan
{t-nakada, kuni}@jaist.ac.jp

Abstract. This paper discusses our algorithm for finding subgroups among personal homepages. Assuming that personal homepages usually carry personal information, we have developed an algorithm that allows us to automatically find potential patterns from them. For example, when the algorithm is applied to personal homepages at some school, we can approximate the ratio between the number of students interested in information science and that of students interested in social science. In the experiment, we successfully created subgroups that showed characteristics of the school. Also, we found relations between subgroups that are important for enhancing human activity.

1 Introduction

The study of relations between the real world and the cyberspace has recently received much attention from researchers. The relations are such that the real world is influenced by cyberspace, and vice versa. CommunityWare, e.g., [2], [8] is a research based on the former relation. The purpose of the researches is mainly to enhance human activities by using cyberspace. On the other hand, the purpose of the latter relation is mainly to understand real world phenomena through the cyberspace. For example, in order to understand user behavior, e-Mail, News, Web server log, and so on have been analyzed, e.g., [1], [7]. Our study belongs to the later type of research; the purpose is to develop an unsupervised learning algorithm for finding interesting patterns from personal homepages. The algorithm finds potential relations between personal homepages by finding subgroups of them.

In other words, our algorithm finds subgroups among personal homepages that are sets of web documents. Web document grouping has been applied to the field of information retrieval to achieve better efficiency and to smmarize results from search engines. Our study is different from these in two ways. First, we are finding subgroup for knowledge discovery. Second, instead of web pages, we treat web sites, sets of web pages, as our input data so that resulting subgroups carry additional information, i.e., owners of personal homepages.

This paper is organized as follows: Section 2 describes our proposed algorithm for finding subgroups among personal homepages. Section 3 goes over two experiments we conducted and provides an evaluation of our algorithm. Conclusions will be given in Section 4.

G. Grieser et al. (Eds.): DS 2003, LNAI 2843, pp. 385–392, 2003.

2 Finding Subgroups among Personal Homepages

The input of our algorithm is a set of top pages of personal homepages. In our study, a site is defined as a set of web pages that consist of one top page and the rest located under the top page directory. For example, http://www.jaist.ac.jp/˜t-nakada/ is the top page of the first author's homepage. Similarly, http://www.jaist.ac.jp/˜t-nakada/myself.html is a page that belongs to author's homepage while http://www.jaist.ac.jp/ is a page that does not belong to the author's homepage.

The output of our algorithm is a set of subgroups of personal homepages and a list of keywords that describe each of them.

2.1 Gathering Words and Hyperlinks from Personal Homepages

Our system takes a set of top pages of personal homepages as its input and starts processing by gathering all words and hyperlinks from all pages underneath. It travels from the top personal page recursively through subpages to pick up all hyperlinks and words that show up. The words are then transformed by a light stemming algorithm (deleting word prefixes and suffixes and reducing plural to singular), non-nouns are removed, and non-word tokens (such as numbers, HTML tags and most punctuation) are stripped. Words and hyperlinks that occur only in a single personal homepage are ignored to reduce time to compute.

This hyperlink and word information is organized into the data structure seen in Fig. 1. On the left hand side lies a list of personal homepages, and on the right hand side lies a list of all words and hyperlinks from personal homepages. A pointing arrow between the left side and the right side means that the personal homepage at the arrowtail has the content at the arrowhead.

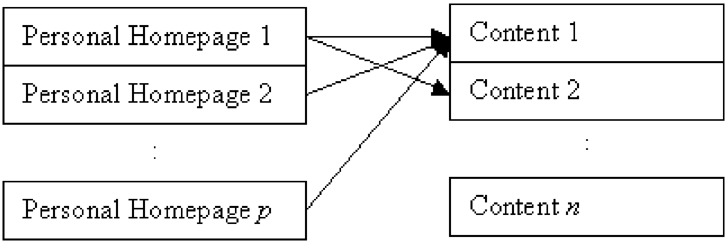

Fig. 1. Data Structure of content information. Contents are either words or hyperlinks in personal homepages

2.2 Finding Subgroups Algorithm

The following shows the algorithm that is applied to the list of contents on the right hand side in Fig. 1 to construct subgroups. The bold strings are to be explained later.

1. Choose a seed content from the list of contents based on **the criterion discussed in 2.3**
2. From the list of contents, get contents similar to the seed content (**similarity measurement discussed in 2.4**)
3. Construct a subgroup from the seed content and contents from 2
4. Iterate 1-3 until obtaining an expected number of sucgroups

We need to repeat the following steps as many times as the number of subgroups desired. The time taken to compute this algorithm is O(mn) where m is the number of subgroups to be created and n is the number of contents.

Quering search engine is used in the procedure 1, 2 (the detail is discussed later). Because our algorithm depends on it, although computational cost is O(mn), our algorithm is very slow in the real time.

2.3 Criterion for Selecting Seed Content

We are to select our criterion for selecting a seed content depending on the purpose of finding subgroups. For example, if the purpose is to find the most predominant contents, the criterion should be to select a content that is used in a large number of personal homepages. However, the result may be trivial because, in the case of hyperlinks as contents, everyone already knows the most famous hyperlink such as http://www.yahoo.com/, so consequent subgroups are not much interesting, or you cannot expect the algorithm to find characteristics of the input data.

Therefore, we developed the following criterion in order to find subgroups from contents famous in a given domain but not in general. We use a score measurement denoted by *score* for a given $content_i$ and chose such a content that *score* in definition 1 is the largest.

Definition 1. $score(content_i)$ for selecting a seed content is defined as the score in the domain ($d\_score(content_i)$) minus the score in general ($g\_score(content_i)$).

$$score(content_i) = d\_score(content_i) - g\_score(content_i) \tag{1}$$

$$d\_score(content_i) = \sum_{n \in Personal\ Homepages\ that\ have\ content_i} content\_score(n) \tag{2}$$

$$content\_score(n) = 1/\#\, of\ All\, Contents\, in\, the\, Personal\, Homepage\, n \tag{3}$$

$$g\_score(content_i) = \#\, of\ Pages\, that\, have\, content_i\, got\, from\, Search\, Engine \tag{4}$$

where, in equation (1), $d\_score(content_i)$ and $g\_score(content_i)$ are standardized to mean 0 and standard deviation 1 in order to compare two types of score.

Equation (3) produces the effect of non-dependence of personal homepages that have a huge number of pages on obtained subgroups.

2.4 Similarity Measure between Two Contents

We use a similarity measure introduced in REFERRAL [3] and [4]. In these systems, similarity between hyperlinks has the following definition, which we extend to measure similarity between words and similarity between a hyperlink and a word.

Definition 2. Similarity between $content_i$ and $content_j$ is defined by Jaccard coefficient [6].

$$\begin{aligned} & similarity(content_i, content_j) \\ & \quad = \frac{\#\, of\, pages\, that\, have\, content_i\, and\, content_j}{\#\, of\, pages\, that\, have\, content_i + \#\, of\, pages\, that\, have\, content_j} \end{aligned} \tag{5}$$

In equation (5), we used a search engine such as Infoseek(http://www.infoseek.co.jp/) in order to count the number of pages that have a given content as opposed to counting the number of such pages within the target domain since the precision of $similarity(content_i, content_j)$ depends on the size of data.

$similarity(content_i, content_j)$ is to be compared with a cut-off point above which the two contents are regarded as similar and below which they are regarded as not similar. To determine our cut-off point, we performed an experiment where the authors picked up 200 pairs of contents and determined whether they are similar or not. The result suggested that the cut-off point should be 0.04.

3 Experiment and Evaluation

We randomly picked up two hundred personal homepages under Stanford University's official homepage (http://www.stanford.edu/leland/dir.html) and ran our algorithm to find subgroups of them. Table 1 shows the eleven subgroups we discovered. Our algorithm stopped searching subgroups when it found eleven seed contents because there were only eleven seed contents in personal homepages at Stanford University. The first column shows the number of personal homepages in each subgroup, the second column shows the number of contents (words or hyperlinks), and the third column shows the contents. The first content in the contents field is the seed content in particular. Contents with a http:// prefix is a hyperlink, and the others are word contents.

In the first subgroup, the characteristic of our algorithm is conspicuous in that both word and hyperlink contents describe the subgroup. However, the rest of the subgroups turned out to have only one type of contents. The reason is that the value of similarity measure (definition 2) becomes closer to zero when there is a big difference between the two numbers of pages. Even when the two contents are actually similar to the eyes of human, they are determined as not similar by the similarity measure. Also, generally speaking, the number of pages with hyperlink contents is much smaller than the number of pages with word contents. Thus, the algorithm tends to construct more subgroups that can be described only by word contents or hyperlink contents, and not both.

Table 1. Results from personal homepages at Stanford University. This list is arranged in the order of $score(content_i)$ (definition 1) where $content_i$ is a seed content.

Subgroup #	# of Personal Homepage	# of Contents	Contents
1	160	7	stanford, http://www.stanford.edu, harvard, cornell, berkeley, yale, princeton
2	67	1	webauth
3	77	4	apache, index, perl, linux
4	94	22	university, science, student, college, research, faculty, institute, school, department, professor, ...
5	43	7	instructor, classroom, undergraduate, curriculum, enrollment, lecture, exam
6	8	1	coworker
7	9	2	alta, vista
8	11	2	http://www.altavista.com, http://www.google.com
9	15	8	yay, ork, gander, xia, mso, csg, cali, cuz
10	24	7	humanity, anthropology, literary, psychology, interdisciplinary, scholar, discipline
11	5	1	psychologist

Looking at the second column, we are able to see the size of each subgroup at the university as a whole. For instance, seventy-seven people belong to the information science subgroup (the third subgroup from top), and this is about three times as big as the size of the humanities and social science subgroup (the tenth subgroup). Although this may not reflect the actual situation at the university since we can expect that a larger percentage of people from information science own homepages than people from humanities and social sciences, the algorithm can still help us have some idea of what the large entity looks like.

Table 2 shows the other result from MIT. In the same way as Stanford University, we randomly picked up two hundred personal homepages under MIT's official homepage (http://www.mit.edu/Home-byUser.html).

We see some differences between Stanford University and MIT. For example, the humanities and social science subgroup appeared only at Stanford University (the tenth subgroup in Table 1) while the subgroup described by brain, disease, and so on appeared only at MIT.

Fig. 2 is our visualization tool that makes clear two types of relations. One is the relation between individuals (i.e., people in a subgroup have relations between each other in the viewpoint of sharing same interests). The other is the relation between subgroups. The large rectangle represents a subgroup and the small rectangle in a subgroup represents a personal homepage. The size of the large rectangle is proportional to the number of its members. All subgroups are

Table 2. Results from personal homepages at MIT. This list is arranged in the order of $score(content_i)$ (definition 1) where $content_i$ is a seed content.

Subgroup #	# of Personal Homepage	# of Contents	Contents
1	125	1	mit
2	74	2	http://web.mit.edu, http://www.mit.edu
3	99	111	engineer, assistant, guitar, component, keyboard, scientist, genre, installation, contract, ...
4	112	31	research, institute, science, university, analysis, laboratory, publication, analyst, researcher, ...
5	82	30	cambridge, oxford, boston, massachusetts, vienna, netherlands, denmark, greece, austria, hungary, ...
6	43	3	design, engine, career
7	108	87	class, function, level, object, process, value, cource, context, example, java, bulk, solution, ...
8	20	1	mechanic
9	30	4	pdf, acrobat, adobe, format
10	44	26	brain, disease, disorder, blood, patient, tissue, ear, diagnosis, cell, cancer, muscle, pain, symptom, ...
11	25	3	http://www.yahoo.com, yahoo, japan
12	27	1	Photo
13	52	21	biology, chemistry, ecology, physics, mathematics, taxonomy, biochemistry, species, medicine, ...
14	4	2	watt, transmitter

shown in the right panel. When a subgroup is selected, it will appear in the left main panel.

We found relations among subgroups from the result of Stanford University. #10 subgroup can be seen as the connection between #3 and #11 subgroup, and in this case ten people become potential key people who have ability to make interaction between #3 and #11 subgroup in the viewpoint of having the same interests. We think it is important to find potential key people in order to enhance human activity.

4 Conclusion

We discussed our algorithm for finding subgroups among personal homepages. In the Experiment, we successfully created subgroups described by both types of contents while, at the same time, we made clear some issues to be solved in the future.

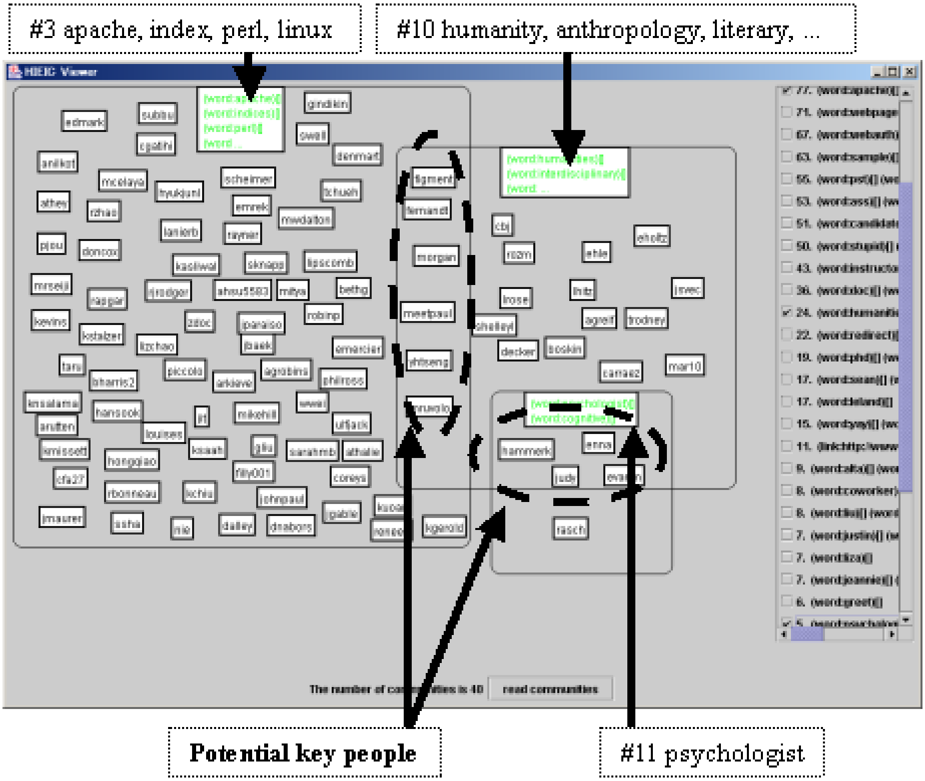

Fig. 2. Sample relations between subgroups and between individuals at Stanford University

We think that one of the applications using our algorithm is to enhance human activity. It is possible to construct a new real world community and produce new interactions between people if subgroups of personal homepages can be found because although the subgroup is not a real world community yet, people in a subgroup share some interest. Other application is to understand human dynamics. Although update frequency of personal homepages is uneven, personal homepages are changed byowners. It is possible to know human dynamics if our algorithm is applied periodically to pick up any change.

References

1. Judith S. Donath, Visual Who: Animating the affinities and activities of an electronic community, in ACM Multimedia 95, 1995.
2. Toru Ishida, Towards Communityware, New Generation Computing, 16(1), 1998, pp. 5–22.
3. H. Kauts, B. Selman, and M. Shah, The Hidden Web, AI Magazine, vol. 18, no. 2, pp. 27–36, 1997

4. Tsuyoshi Murata, Discovery of the Structures of Web Communities, JSAI SIG-KBS-A002-2, pp. 7–12, 2000 (in Japanese)
5. Toyohisa Nakada, Tu Bao Ho and Susumu Kunifuji, Finding Potential Human Communities From Personal Homepages, In Proceedings of the IASTED International Conference ACI2002, pp. 191–196, 2002
6. G. Salton, Automatic Text Processing, Addison-Wesley Publishing Company, Reading, MA, 1989
7. Myra Spiliopoulou, Lukas C. Faulstich, Karsten Winkler, A Data Miner analyzing the Navigational Behavior of Web, in Proceedings of the Workshop on Machine Learning in User Modeling of the ACAI 99, 1999.
8. Yasuyuki Sumi, Kenji Mase, Supporting Awareness of Shared Interests and Experiences in Community, in Proceedings of the International Workshop on Awareness and the WWW, held at the ACM CSCW'2000 Conference, 2000.