Thematic and Temporal Analysis of Thai Web Communities

Kulwadee Somboonviwat† Shinji Suzuki‡ Masashi Toyoda‡ Masaru Kitsuregawa‡

Graduate School of Information Science and Technology, The University of Tokyo†

Institute of Industrial Science, The University of Tokyo‡

E-mail: {kulwadee, suzuki, toyoda, kitsure}@tkl.iis.u-tokyo.ac.jp

ABSTRACT

A web community can be defined as a set of web pages related to a specific topic. Identification of web communities can be done by extracting distinctive structures from a graphical representation of the Web. In this paper we investigate thematic and temporal properties of Thai web communities extracted from two Thai web snapshots (crawled during July 2004 and May 2007).

1. Identification of Thai Web Community

In this paper we use two Thai web snapshots to study the properties of Thai web communities. The first snapshot of the Thai web was crawled during July 2004 using a breadth-first-search strategy and three selected popular Thai web portals as the start seeds. The second snapshot was crawled during May 2007 using a language specific web crawling strategy [2, 3]. The numbers of crawled web pages are 18,344,127 and 1,200,819 pages for July2004 and May2007 datasets respectively.

Most studies of web communities are based on the notion of *authorities* and *hubs* proposed by Kleinberg [4]. An authority is a page with good contents on a topic, and is pointed to by many good hub pages. A hub is a page with a list of hyperlinks to valuable pages on the topic, i.e. the hub page is pointing to many good authorities. A web community is defined as a set of authorities and hubs with a distinctive bipartite graph structure (see Figure 1). We have applied the Companion- algorithm [1] to both Thai web snapshots.

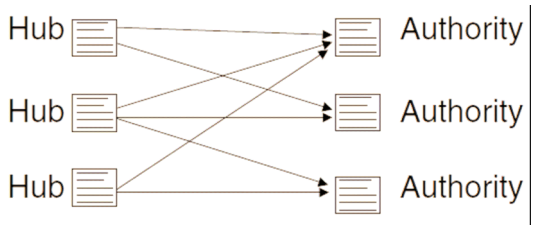

Figure 1: Typical graph structure of hubs and authorities.

Given a seed set, Companion- first builds a subgraph of the Web around the seed, and extracts authorities and hubs in the graph which will then be returned as the output web communities. Based on this technique, we were able to identify 67,752 and 45,436 web communities from July2004 and May2007 datasets, respectively.

Figure 2 shows the log-log plots of size distributions of the web communities found in our datasets. The sizes of the communities (i.e. the number of URLs in a community) are ranging from the smallest community consisting of 2 URLs to the largest community consisting of about 1,700 URLs (July2004) and 130 URLs (May2007). On average, a community will consist of 6 URLs (July2004) and 5 URLs (May2007). As can be seen from Figure 2, the web community size distributions of both datasets on the log-log plots are in the shape of straight line (with a concave in the first part in both plots). Consequently, it follows that the web community size distribution exhibits a power-law distribution. This phenomenon may be explained by a *preferential attachment* (or *the rich get richer* model) where web communities gain new URLs in proportion to how many URLs they already have.

† 東京大学大学院情報理工学系研究科

‡ 東京大学生産技術研究所 喜連川・豊田研究室

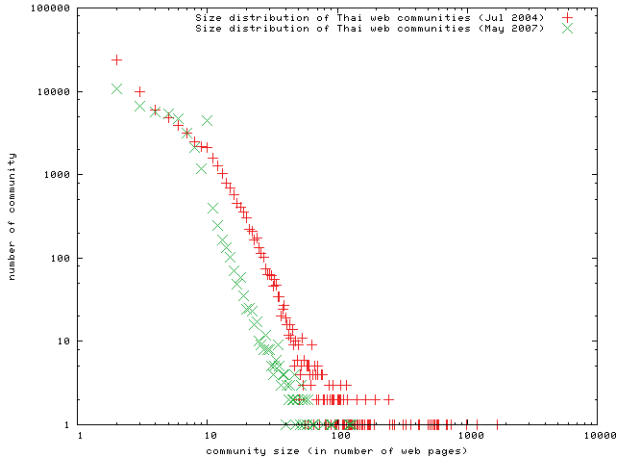

Figure 2: Distributions of the size of Thai web communities adhere to power-law.

By manual inspection, we found that most of large communities are those related to spam and adult Web sites.

2. Analysis of Thai Web Community

2.1 Similarity with Thai Web Directory

We use the following similarity measure to approximate the similarity of the resulting web communities with a web directory:

$$SIM(C_i, D_i) = | C_i \cap D_i | / | C_i | \quad (1)$$

where, C_i is a community and D_i is a category in a web directory. By sampling some web communities from each dataset and applying Equation (1) manually, the calculated similarities with a Thai Web directory (URL at http://www.truehits.net/) are more than 0.95 for July2004 dataset.

2.2 Evolution of Thai Web Community: case study of a Thai university community

We can think of a web community as a representation of a real-world topic, concern, business in the Internet. Thus, it follows that it is possible to study the evolution of certain topics in the real-world society by tracking how topics emerged and evolved in the Web. As a case study of the evolution of Thai web community, in Figure 3, we have show the content of Thai university community in July 2004 (Figure 3(a)) and May 2007 (Figure 3(b)).

(a) July 2004

http://www.kku.ac.th
http://www.mahidol.ac.th
http://www.psu.ac.th
http://www.ru.ac.th
http://www.su.ac.th
http://www.tu.ac.th
http://www.ku.ac.th
http://www.nu.ac.th
http://www.swu.ac.th
http://www.stou.ac.th
http://www.chula.ac.th
http://www.buu.ac.th
http://www.mju.ac.th
http://www.msu.ac.th
http://www.sut.ac.th
http://www.ubu.ac.th

(b) May 2007

http://www.chula.ac.th:80
http://www.ku.ac.th:80
http://www.mahidol.ac.th:80
http://www.ru.ac.th:80
http://www.stou.ac.th:80
http://www.su.ac.th:80
http://www.tu.ac.th:80
http://www.kku.ac.th:80
http://www.swu.ac.th:80
http://www.kmutt.ac.th:80
http://www.psu.ac.th:80
http://www.ubu.ac.th:80
http://www.buu.ac.th:80
http://www.chiangmai.ac.th:80
http://www.mju.ac.th:80
http://www.nu.ac.th:80
http://www.wu.ac.th:80
http://www.msu.ac.th:80

Figure 3: Community of Thai Universities in (a) July 2004 and (b) May 2007.

The number of university in the community has increased from 16 in July 2004 to 18 in May 2007, a slow growth in size. We also found another smaller Thai university which has another kind of evolution i.e. split. This evolution is caused by privatization of some universities in the community. Due to space limit, we cannot show the picture of this split evolution.

3. Conclusion

In this paper we have extracted communities from two Thai web snapshots. By plotting the community size distribution, we have revealed that it fits with the power-law distribution. As a result, the evolution of the size of web communities may potentially be explained by the "rich get richer model". In addition, we also found the existence of spam communities with large size. Then, we showed a high similarity between extracted web communities and a Thai web directory. Finally, we showed an example of the evolution of a Thai university community.

Reference

[1] M. Toyoda, and M. Kitsuregawa. Creating a Web Community Chart for Navigating Related Communities. in *Proc. of 12th ACM Conference on Hypertext and Hypermedia (Hypertext 2001)*, pp. 103-112.

[2] T. Tamura, K. Somboonviwat, and M. Kitsuregawa. A method for language-specific Web crawling and its evaluation. *Systems and Computers in Japan*, 38(2) pp. 10-20, 2007.

[3] K. Somboonviwat, T. Tamura, and M. Kitsuregawa. Finding Thai Web Pages in Foreign Web Spaces. *ICDE Workshops 2006*: 135.

[4] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. in *Proc. of the ACM-SIAM Symposium on Discrete Algorithms,* 1998.