# Discovering Topics to Enhance Communities' Creation from *Links to the Future*

Yukio Ohsawa
TOREST, Japan Science and Technology Corporation (JST)
GSSM, Univ. of Tsukuba
Tokyo 112-0012 Japan
telephone :+81-3-3942-7141
osawa@gssm.otsuka.tsukuba.ac.jp

Naohiro Matsumura
TOREST, Japan Science and Technology Corporation (JST)
Dept. Electronical Eng., Univ. of Tokyo, Tokyo 113-8656 Japan
+81-3-5841-6755
matumura@miv.t.u-tokyo.ac.jp

Mitsuru Ishizuka
Dept. Electronical Eng., Univ. of Tokyo, Tokyo 113-8656 Japan
+81-3-5841-6347
ishizuka@miv.t.u-tokyo.ac.jp

**ABSTRACT** The World Wide Web is a great source of new topics significant for trend birth/creation. Here we propose a method for discovering such topics from the web. The obtained web pages absorb attentions of people from multiple interest-communities, to enforce the spread of latent interest trends. Topics in such pages can be triggers for personal/social progress of interests, beyond the bounds of existing communities.

**Keywords: Community, Latent Trend, Topics**

## 1. INTRODUCTION

Topics sometimes grow into trends absorbing attentions of people. For example, the robot succor games RoboCup seemed just curious when it appeared first, but matched with latent interests of research-communities, i.e. multi-agents, game simulations, robotics, etc. People from these communities, gathering and seeing each other to discuss in the same context was a novel or rare experience. Following such examples, we present a method "Links to the Future" for discovering web pages of new topics to attract multiple communities. Directions to future creation beneficial for communities can be found, as the effect of the topics obtained and visualized by the system.

Studies have been devoted to extracting communities from hyperlinks [1-5]. The method in [3] obtained *hub* pages, linking to *authority*-(linked from many pages as ones obtained by Google [4]) pages popular to established communities. [5] obtained *cores*, groups of densely linked authorities and hubs, for finding emerging communities. On the other hand, our aim is to find premature topics possible to be the seeds of communities not emerging yet. This is for grasping significant yet latent trends, hard to find due to the premature prevalence.

## 2. *Links to the Future* - THE PHILOSOPHY

At least two obstacles exist for *predicting* future trends. The first is the extreme rareness of trend-outbreak signs. If a group of communities often see each other, they would have already evolved into a super-community. Prediction methods relying on past frequent patterns [6] or relatively rare patterns in rich past data [7,8], are not applicable here. The second is the hidden causes for a trend outbreak, hard to be fully considered as features (data attributes) in data analysis/mining. Corresponding to each obstacle, we have two principles for a trend outbreak (see the RoboCup example again):

*Principle 1:* Popular communities exist, each made of people sharing some popular (authorized/established) interests.

*Principle 2:* If a new topic attracts different-interest popular communities, it grows into a fashion among those communities.

## 3. The Method of *Links to the Future*

Corresponding to *principle* 1, interest-sharing communities represented by authority-pages or top pages of Google are obtained by looking at links to those pages [5]. In each community, we regard the highest-ranked page according to Google as the *archive*-page representing the (popular or emerging in the sense of [5]) community. Selecting a single page from one community here is for a comprehensible visualization of the output as in Fig. 1.

Then, corresponding to *principle* 2, pages linked from multiple archive-pages but are not in any community are taken as novel topics attracting multiple communities, called *agora-topic* pages after ancient Egyptian inter-community meetings. That is, an agora-topic is expected to grow into a major concern of a new society greater than existing community obtained above. If people become aware of such an inter-community value of the topic by having people from various communities meet, their communication can lead to an outbreak of the topic. The algorithm outline for obtaining agora-topics is as follows.

**Step 1:** A query representing the interest domain is entered to a search engine (Google here, obtaining $10^5$ to $10^6$ pages).

**Step 2:** Communities, of pages obtained in Step 1, are obtained as in [5] and archive-pages are selected from communities.

**Step 3:** Pages, not in the communities but linked from multiple archive-pages, are obtained as agora-pages.

After the steps here, archive-pages (black nodes), agora-pages (red nodes) and the links between them are visualized as in Fig.1.

## 4. EVALUATION

**Stage 1.** An interest domain is fixed, 4 to 5 (appropriate number for talking) people relevant to the domain gather, and the domain-name is input as a query (e.g. "information retrieval").

**Stage 2.** The output graph adding real and fake red nodes, as if they all were the obtained as agora-pages, is shown to the subjects. That is, some red nodes, not really obtained, were added with red links to black archive-nodes. Subjects reported individual impressions and group-wise ideas in discussions.

We conducted evaluations of the two stages above, for various queries. For query "information retrieval", five subjects engaged in various studies relevant to information retrieval were tested, varying from master-course students to research associates. In stage 2, all the 34 black nodes were said to be strongly relevant to all users' backgrounds. On the other hand, all subjects said the red (agora) nodes showed directions of the area on which they might make some decisions for future work.

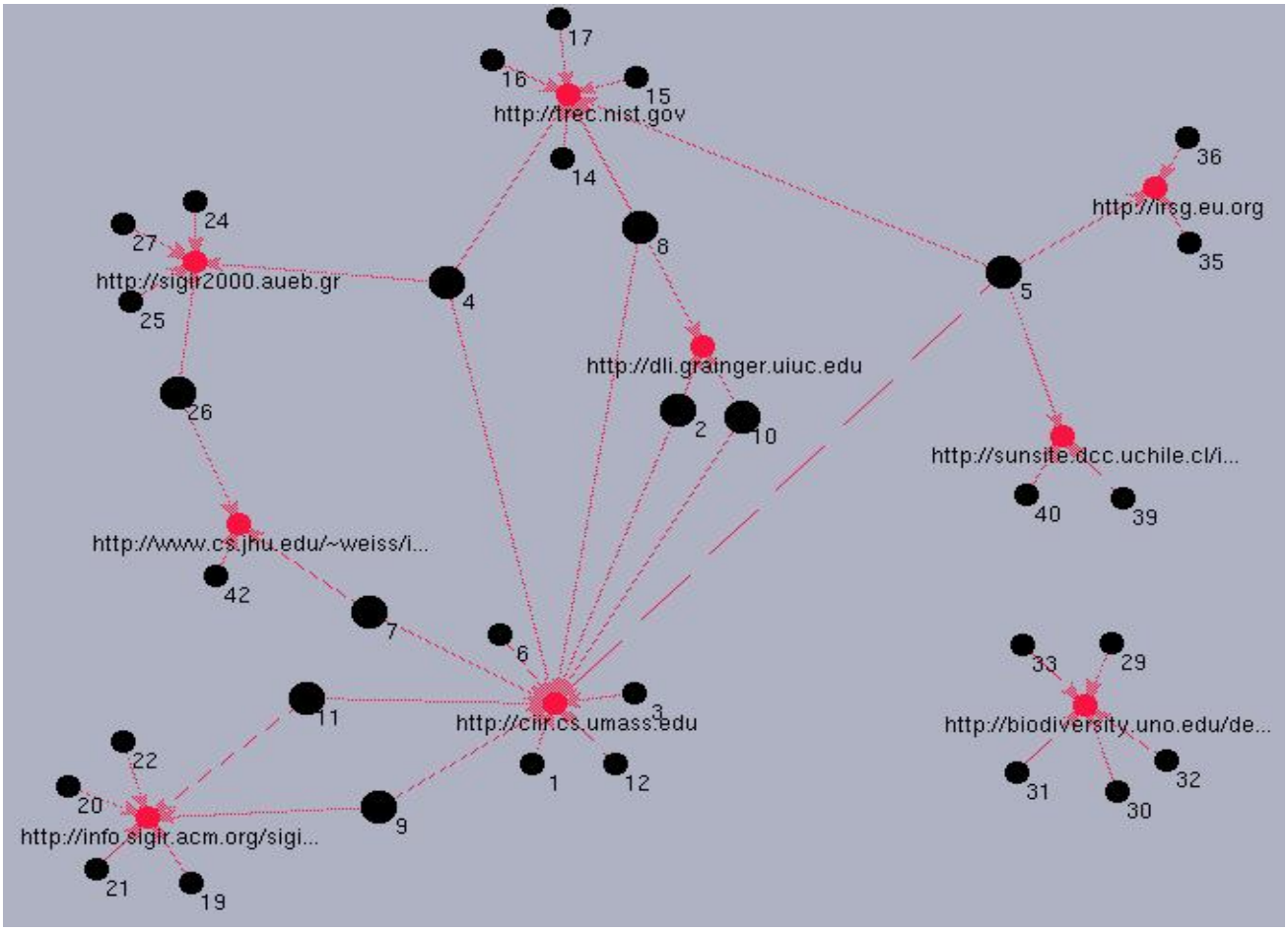


**Fig.1 The output of Links to the Future, for domain query "information retrieval"**

For example, TREC (http://trec.nist.gov/) in a red node appealed to their challenging desire to evaluate their original works on a standard target document collection as dealt in the TREC series. When they read another red page (http://biodiversity.uno.edu/delta), they thought about starting an inter-desciplinary project as DELTA (DEscription Language for Taxonomy - a system for computer processing, including information retrieval, of biological taxonomic descriptions). In fact, DELTA attracts various of researchers - the discussion about its development is increasing the variety of discussants as in the mailing list in http://listserv.surfnet.nl/archives/delta-l.html. Black nodes surrounding these red nodes varied across various sub-communities of information retrieval, e.g., biology and information retrieval people in the case of DELTA.

For all queries including other domains, we had 154 black nodes in total, of which 103 were said to be of established interests in the area. Only 7 black nodes appealed as directions for new decisions for future research. On the other hand, the 37 red nodes included 23 "interesting for planning future work" in subjects' individual impressions, with embodied comments about possible decisions they may make. Their discussions lead to awareness on significant new problems, e.g., "what should the quality evaluation of a search engine be, for measuring user's real satisfactions?" and "who seeks information retrieval techniques for tasks?" through discussing looking at red pages.

## 5. CONCLUSION

This paper presented a method for showing web pages to stimulate various communities to communicate and create new interest trends. A human tends to make decisions affected by belonging communities, especially if the decision is novel and one needs to be somehow encouraged for deciding. The presented method will realize an effective promotion for encouraging people to make such novel activities.

## REFERENCES


[1] Chakrabarti, S. et al, Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. *In Proc. of WWW7* (1998)

[2] Gibson, D., Kleinberg, J. and Raghavan, P. Inferring Web communities from link topology. *In Proc. of 9th ACM Conference on Hypertext and Hypermedia* (1998)

[3] Kleinberg, J. Authoritative sources in a hyperlinked environment, *IBM Research Report RJ* 10076 (1997)

[4] Brin, S. and Page,L. The anatomy of a large scale hypertextual web search engine. *In Proc. of* 7th World-Wide Web conference (WWW7), (1998)

[5] Kumar,S.R., et al, Trawling the Web for Emerging Cyber-communities *In Proc. of WWW8* (1999)

[6] Mannila, H, et al, "Disocvering Frequent Episodes in Event Sequences," in Proc. First Conf. on Knowledge Discovery and Data Mining (KDD95), 1995.

[7] Weiss, G.M. and Hirsh,H. "Learning to Predict Rare Events in Event Sequences," Proc. of Knowledge Discovery and Data Mining (KDD-98), 359-363, 1998.

[8] Suzuki, E. and Kodratoff, Discovery of Surprizing Exception Rules Based on Intensity of Implication, in Principles of Data Mining and Knowledge Discovery, LNAI 1510, 10--18, Springer, 1998.