Discovering Seeds of New Interest Spread from Premature Pages Cited by Multiple Communities

Naohiro Matsumura[1 3 *] Yukio Ohsawa[1 2 **] and Mitsuru Ishizuka[3]

[1] TOREST, Japan Science and Technology Corporation
[2] Graduate School of Systems Management, University of Tsukuba, 3-29-1 Otsuka,
Bunkyo-ku, Tokyo 112-0012, Japan
[3] Dept. of Electronic Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-8656, Japan

Abstract. The World Wide Web is a great source of new topics significant for trend birth and creation. In this paper, we propose a method for discovering topics, which stimulate communities of people into earnest communications on the topics' meaning, and grow into a trend of popular interest. Here, the obtained are web pages which absorb attentions of people from multiple interest-communities. It is shown by a experiments to a small group of people, that topics in such pages can trigger the growth of peoples' interests, beyond the bounds of existing communities.

Keywords: Communities, Interest Spread, Web Links

1 Introduction : Which Topics Grow into Trends ?

Some new topics grow into a prevalent concept, if they satisfy the desire of people for information. Simple and sensational imformation comes quickly spread to fit to uncertain minds desiring information [1].

Our aim is to detect topics which can be spread to satisfy a wide range of people. People first become aware of the topic, understand and accept it. Then, the topic grows to be established and will not decay as easily as lies. Topics leading to such a prevalence is worthwhile finding for commercial/personal benefits. In this paper, the problems addressed are:

1) How and what kind of topics grow into a trendy interest ?
2) How can we support the awareness of human community on such topics ?

In section 2, problem 1) will be discussed on for making a strategy for challenging problem 2). In section 3, we point out why previous Web-mining methods could not find such opinions we aim at. In section 4, a new method *Agora on Links* based on Web links, fitting our goal is presented. The method obtains Web-pages, including topics which are premature but will be accepted as significant

[*] e-mail:matumura@miv.t.u-tokyo.ac.jp
[**] e-mail:osawa@gssm.otsuka.tsukuba.ac.jp

for a wide range of communities, beyond the boundary of personal interests. Conclusional remarks are in Section 6.

2 A Mechanism of the Topic Growth into a Trend

For question 1) above, we go after the mechanism of community extensions. We define a community as a group of people sharing some value, as people in the society of artificial intelligence sharing the value admiring the aim to make AI studies, although their interests are not precisely equal. People in different communities do not see each other usually - if they do, we regard them as one community because they already share established values for which to gather.

If multiple communities have an occasion to meet for talking on a newly born opinion or topic, and if a new value appears to be shared there, participants will evaluate each other positively on the value (as the measure to evaluate other people) and find it meaningful to be re-organized into a greater community sharing the new value.

This means that a new encountering of communities, differing in previous interests but sharing a topic, can be a trigger to the innovation of a new strong idea as pointed in [2,3]. Further, the awareness on various relations (difference, similarity, or others) between one's initial belief (constructed in the community one usually belongs to) and new information (which may come from another community) urges the topic to be understood and established in each community.

For example, the robot succor games RoboCup seemed just curious when it appeared first, but matched with latent interests of various research-communities, who gathered and achieved new innovations.

3 The Concept of *Agora on Links*

Following the topic-spread model above, we present a method "Agora on Links" for aiding the discovery of web pages on new topics premature but attracting multiple interest-communities, by visualizing those topics and their links with the interests of already established communities.

Studies have been devoted to extracting communities from hyper-links on the Web ([4]-[8]). The method in [6] obtains *hub* pages, linking to *authority*-pages (linked from many pages as ones obtained by Google [7]) popular to established communities. The method in [8] obtains *cores*, i.e. 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.

At least two obstacles exist for predicting future trends. The first is the extreme rareness of trend-outbreak signs. As pointed in section 2, 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 [9] or relatively (i.e. not very) rare patterns in rich past data [10,11], are not applicable here. The second is the hidden causes for a trend outbreak i.e., the interest- context of each community and the relationships among communities, hard to be fully considered as features (data attributes) to be used in data analysis/mining. In order to cope with these obstacles, we apply the two principles below, and visualize the signs of a trend for stimulaing user's imagination of hidden causes rather than fully automated prediction.

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

Principle 2: If a new topic attracts different-interest popular communities, it will grow into a new trend to grow among those communities.

4 The Method of *Agora on Links*

Corresponding to Principle 1, the set of authority-pages or top pages representing each community, from the output pages from Google are obtained by looking at links to those pages [8]. We regard the page of the highest-rank according to Google, in each community, as the archive-page representing a (popular or emerging in the sense of [8]) community. Selecting a single page for representing one community here is in order to form a comprehensible visualization of the output as shown later in Fig.1.

Then, corresponding to principle 2, pages linked from multiple archive-pages but are not in any community themselves are taken as novel topics attracting multiple communities, called *agora-topic* pages after the name of ancient Egyptian inter-community meetings. The algorithm outline for obtaining Web pages representing agora-topics is as follows.

Step 1: A query representing user's 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 [8] 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. Having all obtained results by here, archive-pages (black nodes), agora-pages (red nodes) and the links between them are visualized as in Fig.1.

5 Evaluation: Agora Topics Spread in Groups of People

The Evaluation Experiment :

Stage 1. An interest domain is fixed, a group (appropriate number for talking) of people relevant to the domain gathered, 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 really 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 exchanged ideas in the group.

We conducted evaluations of the two stages above, for various queries.

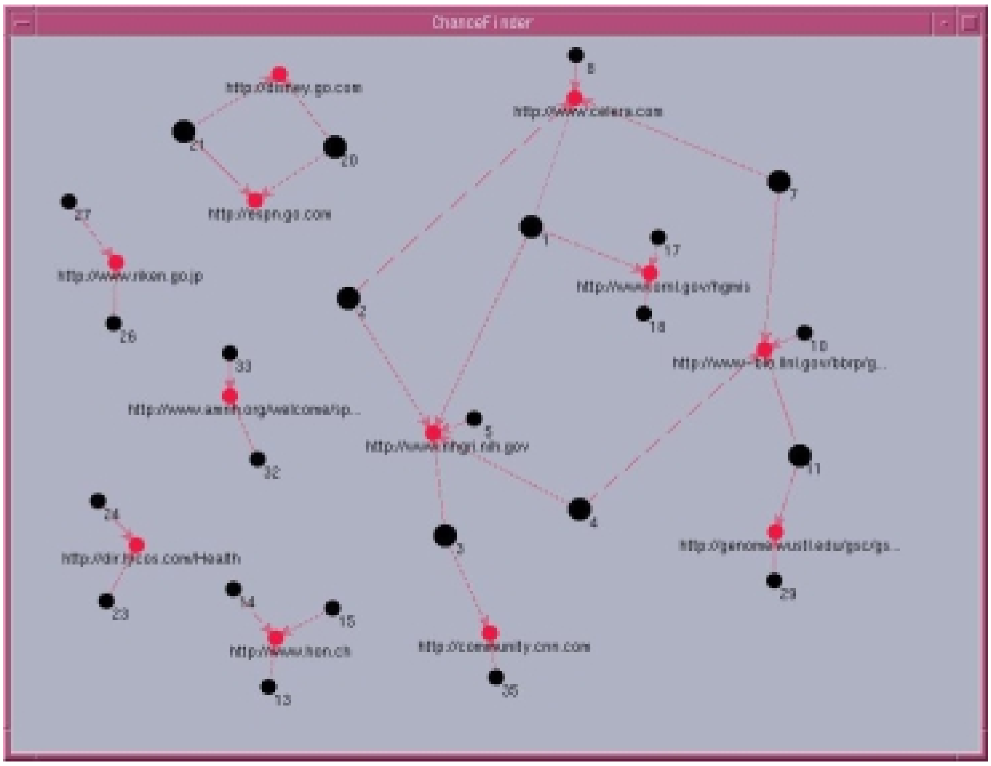

Fig. 1. The output of *Agora on Links*, for domain query "human genome."

For query "human genome," three subjects in Human Genome analysis area were gathered. In stage 2, all Web pages in the black nodes in Fig.1 were said to represent common knowledge in the area backgrounds. On the other hand, subjects said the red (agora) nodes showed pages of growing interests. More concretely, red nodes were said to be the target of interests from various institutes in or relevant to the area, which seldom meet in frequent workshops but are considered to achieve significant discoveries if they exchange knowledge with each other.

For example, in Fig.1, the largest cluster of nodes and links is the group of American human-genome research institutes, looking at growing interest trends, e.g., the venture-company Celera the leading institute NIH in genome research.

On the other hand, the left-hand small cluster is Japanese institutes, looking at (linking to) Riken, the most active in the identification of mouse genomes. In fact, these institutes in red are ones who have data sources of human or mouse genomes, and it is useful for researchers in other institutes to look at those data. In other words, Web pages having such data is the agora for researchers.

For all queries including other domains, we had 53 black nodes in total, of which 37 were said to be of established interests in the area. Only 3 black nodes appealed as directions for new decisions for future research. On the other hand, the 12 red nodes included 8 "interesting for thinking of future work" in subjects' individual impressions, with embodied comments about possible decisions they may make. Their discussions lead to their awareness on significant new problems, e.g., "what should the evaluation criteria of a search engine be, for measuring user's real satisfactions?" and "who seeks information retrieval techniques for tasks?" through discussing with looking at red pages.

6 Conclusions

A hypothesis on the growth of a minor topic to be a wide spread interest to the majority of people is given, and an algorithm is shown for detecting such topics from the Web, corresponding to the model. The method was applied to Web links in specific domains, and it was shown the visual output aids in the discovery and the growth of meritorious concepts.

References

1. Allport,G.W and Postman, L., *The Psychology of Rumor*, H.Holt.(1947)
2. Clark,K.B., and Fujimoto,T. *Product Development Performance: Strategy, Organization, and Management in the World Auto Industry.* Boston: Harvard Business School Press. (1990)
3. Koestler, A., *The Act of Creation*, New Yourk: Liveright (1964)
4. Chakrabarti, S. et al, Automatic Resource Compilation by Analyzing Hyperlink structure and Associated Text. In *Proc. of WWW7* (1998)
5. Gibson, D., Kleinberg, J. and Raghavan, P. Inferring Web communities from line topology. In *Proc. of 9th ACM Conference on Hypertext and Hypermedia* (1998)
6. Kleinberg, J. Authoritative sources in a hyperlinked environment, IBM Research Report RJ 10076 (1997)
7. 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)
8. Kumar,S.R., et al, Trawling the Web for Emerging Cyber-communities In Proc. of *WWW8* (1999)
9. Mannila, H, et al, "Disocvering Frequent Episodes in Event Sequences" in *Proc. First Conf. on Knowledge Discovery and Data Mining (KDD95)*, 1995.
10. Weiss, G.M. and Hirsh,H. "Learning to Predict Rare Events in Event Sequences," in *Proc. of KDD-98*, 359-363, 1998.
11. 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.