Socio-Sense: A System for Analysing the Societal Behavior from Long Term Web Archive*

Masaru Kitsuregawa[1], Takayuki Tamura[1,2],
Masashi Toyoda[1], and Nobuhiro Kaji[1]

[1] Institute of Industrial Science, The University of Tokyo,
4–6–1 Komaba, Meguro-ku, Tokyo, 153-8505 Japan
{kitsure,tamura,toyoda,kaji}@tkl.iis.u-tokyo.ac.jp
[2] Information Technology R&D Center, Mitsubishi Electric Corporation,
5–1–1 Ofuna, Kamakura-shi, Kanagawa, 247–8501 Japan

Abstract. We introduce Socio-Sense Web analysis system. The system applies structural and temporal analysis methods to long term Web archive to obtain insight into the real society. We present an overview of the system and core methods followed by excerpts from case studies on consumer behavior analyses.

1 Introduction

Socio-Sense is a system for analysing the societal behavior based on exhaustive Web information, regarding the Web as a projection of the real world. The Web is inundated with information issued from companies, governments, groups, and individuals, and various events in the real world tend to be reflected on the Web very quickly. Understanding the structure of the cyber space and keeping track of its changes will bring us deep insight into the background and the omen of real phenomena. Such insight cannot be achieved with current search engines, which mainly focus on providing plain facts.

The system has been developed from the ground up in the following directions:

- Web archive consisting of 9 years' worth of Japanese-centric Web contents, which enable long term historical analyses.
- Web structural analysis methods based on graph mining algorithms and natural language processing techniques. By grouping topically related Web pages, one can browse and navigate the cyber space at a macroscopic level. On the other hand, microscopic information such as product reputations can also be identified.
- Web temporal analysis methods to capture events in the cyber space such as emergence, growth, decay, and disappearance of some topic, or split and merger among topics.

* Part of this research was supported by the Comprehensive Development of e-Society Foundation Software program of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Y. Zhang et al. (Eds.): APWeb 2008, LNCS 4976, pp. 1–8, 2008.

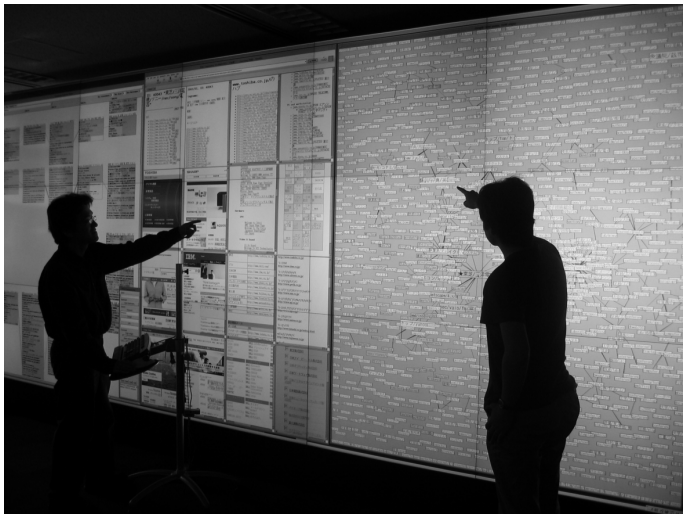

Fig. 1. Display wall at the frontend of the system

The above elements have been integrated into the system to conduct case studies assuming corporate users' needs, such as tracking of reputations of brands or companies, grasping of consumer preferences, and analysis of consumers' lifestyles.

2 System Overview

At the base of the system is the Web archive, which consists of Japanese-centric Web contents[1] and their derivatives accumulated in a bunch of storage devices. The archived contents span 9 years now.

The Web archive started as a mere collection of yearly snapshots obtained from each run of a batch-mode crawler, and has evolved towards a general temporal database, where new versions of each Web page are independently appended. The associated crawler, which keeps running on a bunch of servers, now operates in the continuous mode, estimating update intervals of Web pages to visit them adaptively. As a result, the minimum time resolution between versions has been reduced to a day.

The URL-time index of the Web archive supports tracking of history of a URL, and crosscutting of whole URLs at arbitrary times. Contents of different periods can be uniformly searched with full text queries. Thus, history of ocurrence frequency of specific words can be easily obtained.

Though the Web archive supports exporting of its subset in one of general archive formats such as tar, the system tightly couples the Web archive with an analysis cluster to avoid overhead of moving around huge amount of data. With this parallel scanning mechanism, contents are extracted from the Web

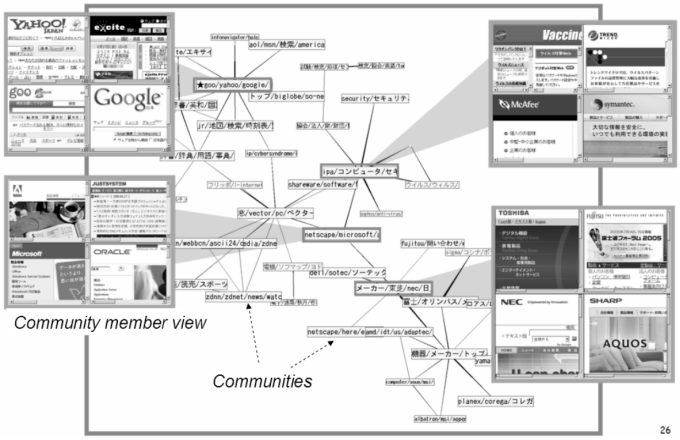

Fig. 2. Community chart on "computers"

archive and dispatched on the fly to one of the cluster nodes, where an instance of application-specific content processing loop is running. The system also takes care of load balancing among the cluster nodes.

The results of the analyses are significantly reduced in size compared with their input, but they tend to be still too complicated to present on space-limited desktop screens. Thus, we built a display wall with 5k x 3k pixels to visualize complex results nicely. Figure 1 shows the display wall showing the results from structural and temporal analyses, which are described next.

3 Web Structural Analysis

Topically related Web pages tend to be connected with relatively large number of hyper-links and reside topologically near in the Web graph. Leveraging this property, we obtained sets of related pages by extracting dense subgraphs from the whole Web space. We call each set of related pages a Web community. Subgraphs dense enough to comprise Web communities are commonly observed in various areas, from home pages of companies in the same category of industry, to personal pages mentioning the same hobbies.

After having extracted the Web communities exhaustively, communities were linked to each other according to the sparse part of the Web graph. This resulted in an associative graph with communities as nodes and degrees of relationship among communities as edges. This high level graph is called a community chart and serves as a map of the cyber space in terms of communities. Figure 2 is a subset of the community chart which relates to a term "computer." Each rectangle represents a community and related communities are connected with links. Member pages are also shown for four communities, namely computer hardware

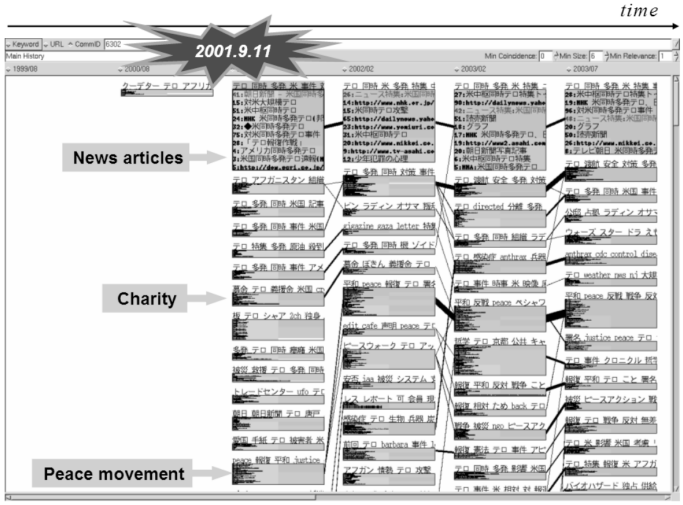

Fig. 3. Evolution of communities on "terror"

vendor community, software vendor community, security information/vendor community, and portal/search engine community (from right-bottom to left-top). This can be regarded as inter-industry relationship exposed on the Web. A graphical frontend is provided to explore the community chart, modifying the visible communities and their layout interactively.

In addition to the relationships among Web pages, it is also important to analyze textual data in the Web pages themselves. One challenge is reputation extraction. The Web text contains consumer-originated reputations of products and companies, which are useful for marketing purpose. However, extracting reputations is not trivial. It requires huge lexicon that exhaustively lists up affective words and phrases, and it is costly or even impractical to build such lexicon by hand. To tackle this problem, we employed linguistic patterns and a statistical measure in order to automatically build the lexicon from the Web archive[3]. Using this lexicon we developed a reputation extraction tool. In this tool, reputations of a query are extracted from the archive and displayed to users. Besides original texts, the number of positive/negative reputations and facets on topic are also presented. These functions provide users brief overview of the result.

4 Web Temporal Analysis

By arranging community charts derived for different times side-by-side, we can track the evolution process of the same communities. Linking communities of different times can be accomplished by regarding the member URL set as the

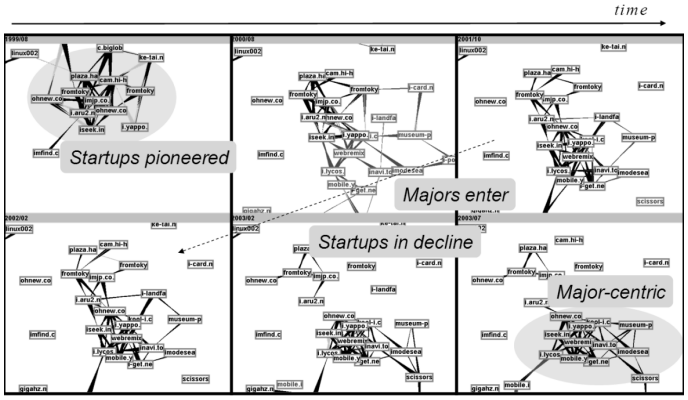

Fig. 4. Structural evolution inside of a community

identity of a community. Some URLs join and leave a community over time, and sometimes communities split and merge. The lack of counterpart of a community implies emergence or disappearance of the community.

Figure 3 shows a screenshot of a tool for visualizing the evolution process of communities. Each column corresponds to different times and each rectangle represents a community with its member URLs shown inside. Inter-community links between adjacent time slots are depicted instead of links within each time slice. This example reveals that right after the September 11 attacks, terror-related communities emerged abruptly. The method can be applied to investigate emeregence of new information, transitions of topics, and sociological phenomena.

The above methods for Web structural and temporal analyses can be combined to visualize the evolution of graph structures themselves[2]. This is most useful for Web graphs at page granularity in that subgraphs not dense enough to form a community can be captured. We can observe the characteristics of graph structures at embryonic stage of community formation and at stage of community growth.

Figure 4 shows evolution of the graph structure inside Japanese mobile search engine communities. Each of 6 panes displays the graph structure at the corresponding time. Each pane is layed out in a "synchronized" manner, where corresponding pages (nodes) are located at similar positions in each pane. We can easily identify what has happened at each stage by interactively manipulating the graphs. At the early stage, search services for mobile phones in Japan were mainly provided by startups or individuals. We can observe that, however, after major companies entered the industry, the center of the community has gradually moved to such companies.

A lot of new words are born and die every day on the Web. It is interesting to observe and analyze dynamics of new words from linguistic perspective. To analyze the dynamics of words, we estimate the frequency of new words in each

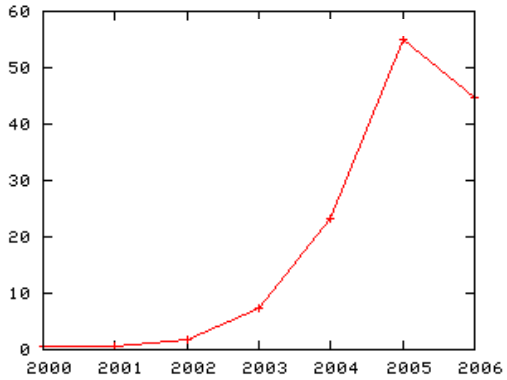

Fig. 5. Evolution of new verb *gugu-ru* (google in Japanese)

year. Because Japanese does not have word separator and it is often difficult for conventional technique to accurately estimate the freqnency of new words, we employed Support Vector Machine (SVM) to extract new verbs and adjectives from the Web. Since verbs and adjectives usually inflect regularly, character n-gram was used as features of SVM.

Figure 5 shows evolution of new verb *gugu-ru* (google in Japanese). The y-axis represents the normalized frequency in the Web archive. We can see that *gugu-ru* has become popular in recent years although it was not frequently used in 1999.

5 Consumer Behavior Analysis

Prevalence of blogs drastically reduced the burden for individuals to express their opinions or impressions, and blogs have been recognized as an influential source for decision making of individuals because blogs have agility and reality in contrast to information originated in companies and mass media. Companies also start recognizing significance of blogs as a tool for grasping consumers' behavior and for communicating with consumers more intimately.

Because of this situation, we applied the methods for Web structural and temporal analyses to analysis of consumer behavior based on blog information[1]. As a consequence, we obtained a visualization tool for inter-blog links. This tool can visualize link structure at arbitrary time, which can be intuitively adjusted with a slide bar. We can easily replay the temporal evolution of link relationships[4].

Considering inter-blog links as an indication of topic diffusion, we can observe how word-of-mouth information has spread out via blogs. For example, we succeeded in identifying a source blog for a book which got drastic popularity gain through WOM. Figure 6 shows temporal changes in links to the source blog (circled). We can observe that, over time, the site gets more and more links.

[1] This part of work was done through a joint effort with Dentsu, Inc. and Senshu University.

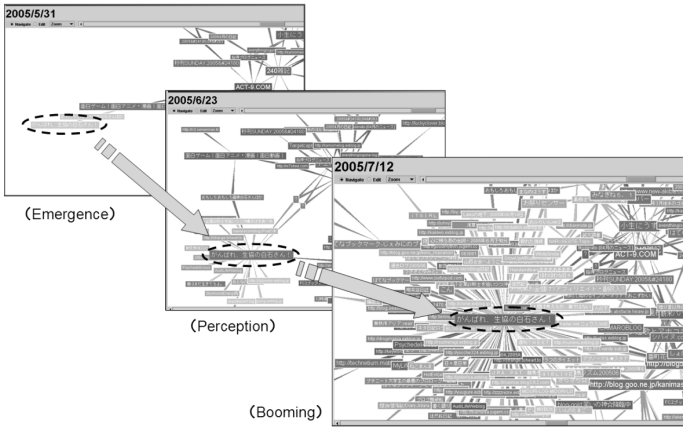

Fig. 6. Popularity evolution of a WOM source

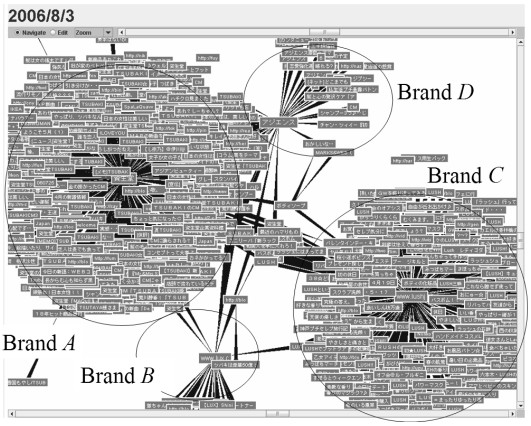

Fig. 7. Comparison of commodity brands

For companies, it's more important to figure out how they and their products are perceived. Figure 7 shows a link structure among blog entries and Web sites of commodity brands in the same industry. Blogs mentioning (linking to) a brand gather around the brand's site. Thus, attractive brands can be easily identified (in this case, brands A and C).

6 Summary

Socio-Sense system was overviewed. The combination of long term Web archive, graph mining algorithms, and natural language processing techniques enabled the system to figure out structure and evolution process of the cyber space.

We have confirmed through various case studies that our system is effective for understanding behavior of the society and people.

References

1. Tamura, T., Somboonviwat, K., Kitsuregawa, M.: A method for language-specific web crawling and its evaluation. Syst. Comp. Jpn. 38, 10–20 (2007)
2. Toyoda, M., Kitsuregawa, M.: A system for visualizing and analyzing the evolution of the web with a time series of graphs. In: Proc. of Hypertext, pp. 151–160 (2005)
3. Kaji, N., Kitsuregawa, M.: Building lexicon for sentiment analysis from massive collection of HTML documents. In: Proc. of EMNLP-CoNLL, pp. 1075–1083 (2007)
4. Toyoda, M., Kitsuregawa, M.: What's really new on the web?: identifying new pages from a series of unstable web snapshots. In: Proc. of WWW, pp. 233–241 (2006)