A Method to Extract Unsteadiness of Concept Attributes Based on Weblog

Yosuke Horiuchi, Osamu Uchida

***Abstract*—** Concept bases are composed of a collection of concept attributes and used for multiple purposes such as improving efficiency of information retrieval and making commonsensical judgments using computers recently. To construct concept bases, the data of the dictionaries is generically used. However, concept attributes are not always static, that is, some of them shift by the influence of various events and incidents. For example, it is to be expected that the attributes of the sports in the concept attribute of the country holding some sports event are stronger than usual time, or they are append to the concept attribute of the country. In this study, we consider the application of weblogs to extract the fluctuations of concept attributes. Many of articles of weblogs are influenced by the news, and the number of documents of weblogs is very large. Then, in this study, we propose a new method to extract the influence of various events and incidents to attributes by regarding the tags given to an article as an attribute of the words in the article, and verify the effectiveness of our method by an experiment.

***Keywords*—** Collective intelligence, Concept attribute, Concept base, Consumer-generated media, Tag, Weblog.

I. Introduction

A concept base is a knowledge database of words, generally consisting of sets of words. These words are attributes representing the expressed concepts (words having meanings and characteristics closely related to referenced words) and attribute values (closeness of the relationship between a word and an attribute) (see, e.g., [1], [2]). For example,

$$\left\{ \begin{array}{lcl} "bird" & = & \{("egg",3),("feather",1),("wing",3),\cdots\} \\ "car" & = & \{("morter",3),("wheel",2),("windoe",1),\cdots\} \\ \vdots & & \vdots \end{array} \right\}.$$

Concept bases are used for multiple purposes such as improving efficiency of information retrieval [3], [4] and making commonsensical judgments using computers [5], [6].

To build a large scale concept base, it is necessary to obtain concepts automatically using computers. Various methods have been proposed for this purpose. A typical method uses entries in a dictionary as concept words and definition sentences of entries as attributes to build a concept base [2], [7]. Additional methods have been proposed to expand concept bases by using newspapers and documents on the web [8]–[10] or by modifying the degrees of relationship by considering unsteadiness of concepts arising from multiplicity of meaning [11]. However, these conventional methods do not account for temporary shifts in concept attributes caused by various events and incidents. Not all concept attributes have been firmly established. For example, it is to be expected that the attributes of the sports in the concept attribute of the country holding some sports event are stronger than usual time, or they are append to the concept attribute of the country. That is,

$$\begin{array}{l} "Japan" = \{("Asia",2),("country",5),("sports",1),\cdots\}, \\ \quad\downarrow \\ "Japan" = \{("Asia",1),("country",2),("sports",5),\cdots\}. \end{array}$$

By reflecting on such shifts in concept attributes, it is expected that the accuracy of information retrieval services (see, e.g., [12]–[15]) and information recommendation systems (see, e.g., [16]–[19]) will be improved.

On the other hand, in recent years, there has been a rapid spread of media such as blogs, wikis, message boards, customer review sites, and social networking sites (SNSs), which make it possible for individuals to generate information more easily. These are collectively referred to as “consumer-generated media” or “user-generated content,” and the numbers of such media users are growing at an explosive rate. For example, Facebook [20], a leading SNS, has more than 600 million users worldwide and was the top visited website in the United States in 2010. The widespread use of these media has made it possible to distribute information without requiring technical knowledge. The number of weblog users continues to increase and a large number of documents are distributed daily on the Internet. Although the contents of weblogs vary, many of them refer to writers’ personal experiences and interests, and thus, weblogs are likely to be influenced by various events and incidents reported at the time of writing. In other words, the impact of an event, accident, or crime is considered to be reflected in the number of weblog articles referring to it.

This study uses weblog articles as a corpus, proposes a method to extract concept attributes that may shift depending on

Manuscript received April 8, 2011.

Y. Horiuchi was with the Graduate School of Engineering, Tokai University, 4-1-1 Kitakaname, Hiratsuka, Kanagawa, 259-1292 Japan. He is now with the Yahoo Japan Corporation, Tokyo, Japan (e-mail: yousuke.horiuchi@gmail.com).

O. Uchida is with the Department of Human and Information Science, Tokai University, 4-1-1 Kitakaname, Hiratsuka, Kanagawa, 259-1292 Japan (corresponding author to provide phone: +81-463-58-1211; fax: +81-463-58-9461; e-mail: o-uchida@tokai.ac.jp).

the subject, and verifies its effectiveness by an experiment.

The remaining sections of this paper are organized as follows. In Section II, we shall introduce the proposed method. In Section III, we shall account for experimental results. In Section IV, we shall summarize results of this study and mention some future research directions.

II. Proposed Method

A. Tags and Weblog Services

In this study, it is assumed that a concept attribute held by a writer of a weblog article and tags (labels) attached to the article are equivalent. In weblogs, tags represent weblog user attributes for each user and are attached to each article to represent the category to which an article belongs. The presence of tags and the manner in which they are attached vary depending on the weblog service. For the purpose of this study, weblog articles were collected from services that allow tags to be attached to each article because it is necessary for tags to reflect topics that differ each day. For example, in Blogger [21] the user can attach tags (labels) for each blog articles (Fig. 1). In this study, the subject of research is restricted to Japanese, then we selected the Yahoo! Japan Blog service [22], because the number of users of the Yahoo! Japan Blog service is very large in Japan, and in which the user can attach tags for each blog articles.

B. Attribute Vector and Concept Base

In this study, we define the attribute vector $\boldsymbol{W}$ of a concept word w_i by

$$\boldsymbol{W}(w_i)=\{(p_{i,1},q_{i,1}),(p_{i,2},q_{i,2}),\cdots,(p_{i,m_i},q_{i,m_i})\},$$

where $p_{i,j}$ is the j th attribute, $q_{i,j}$ is representing the j th attribute's strength of how relevantly it explains the concept word w_i, and m_i is the number of attributes of the concept word w_i. For example,

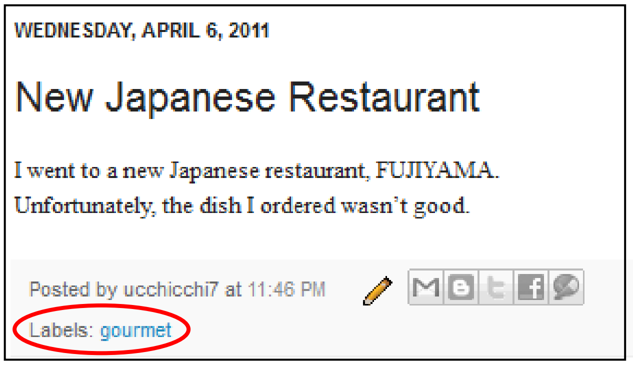
$$\boldsymbol{W}(\text{"}bird\text{"})=\{(\text{"}egg\text{"},2),(\text{"}feather\text{"},1),(\text{"}wing\text{"},3),\cdots\},$$

$$\boldsymbol{W}(\text{"}car\text{"})=\{(\text{"}morter\text{"},3),(\text{"}wheel\text{"},2),(\text{"}window\text{"},1),\cdots\}.$$

The concept base K is a set of attribute vectors [1], [2], that is,

$$K=\{\cdots,\boldsymbol{W}(\text{"}bird\text{"}),\cdots,\boldsymbol{W}(\text{"}car\text{"}),\cdots\}.$$

In this study, we introduce the concept of time into the attribute vectors. These attribute vectors makes it possible for us to make concept base which change according to the day and time. By using this type of concept base, it is expected that the accuracy of information retrieval services and information recommendation systems will be improved. To accomplish the purpose, we use weblog articles as a corpus. We will give an explanation of the proposed method below.



WEDNESDAY, APRIL 6, 2011

New Japanese Restaurant

I went to a new Japanese restaurant, FUJIYAMA.
Unfortunately, the dish I ordered wasn't good.

Posted by ucchicchi7 at 11:46 PM

Labels: gourmet

Fig. 1 Example of blog article with tag (label)

The Yahoo! Japan Blog classifies weblog articles into 15 master categories, 56 intermediate categories, and 351 subcategories. A weblog writer selects a subcategory that he or she thinks is most appropriate for the piece of writing. The selected category is attached to the article as a tag. In our experiment, a total of 47 attributes (Table 1) were defined with reference to the intermediate categories. In this study, concept words were selected from nouns appearing in the collected weblog articles. The attribute vector of a concept word were generated by voting to the attributes to which the tags attached to the weblog article which contains the concept word belong. Given concept words w_i ($i=1,2,\cdots,n$) at time t (here, n is the total number of nouns appearing in the collected weblog articles), the attribute vector of these concept words $\boldsymbol{W}^{(t)}(w_i)$ is defined as follows.

$$\boldsymbol{W}^{(t)}(w_i)=\{(p_1,q_{i,1}^{(t)}),(p_2,q_{i,2}^{(t)}),\cdots,(p_{47},q_{i,47}^{(t)})\}.$$

Here, p_j ($j=1,2,\cdots,47$) represents 47 types of attributes. In addition, $q_{i,j}^{(t)}$ is the attribute value of j th attribute of concept word w_i at time t. In this study, attribute value is defined as the number of appearances of an attribute and is determined by using weblog articles posted during the past 24 hours from time t.

III. Experiment

A. Outline

In this experiment, a large number of weblog articles were collected, and by counting the co-occurrence relationships between words appearing in each article and the tag attached to the article, it was examined whether there were any shifts in concept attributes. The experiment was carried out by extracting nouns from each weblog article by using morphological analysis and these nouns were used as concept words. (A Japanese language morphological analysis API provided by Yahoo! Japan [23] was used.) The weblog articles were collected from 00:00 a.m. on December 3 to 00:00 a.m. on December 26, 2010 by using a weblog article collection program implemented in the PHP language on a Linux server (the collection was interrupted from 00:00 a.m. on December 13 to 00:00 a.m. on December 16 because of maintenance of the collection server). The attribute vector $\boldsymbol{W}^{(t)}(w_i)$ of concept word w_i was determined by using the reference time of 00:00 a.m. each day from December 4 to December 26.

B. Results and Observations

In the experiment, a total of 405,210 weblog articles were collected. Fig. 2 shows the number of blog articles belonging to each category. Fig. 3 shows the time-series of attribute values (normalized frequency) of all 47 attributes in the case where the concept word is "Russia." As shown in this figure, the attributes, for example, "economics," "politics and political activity," "international affairs," and "sports" are unsteadiness. Although both concept words for which the frequency of appearance of attributes changed and those for which frequency of appearance of attributes was stable were present, here, observations are presented by using time-series graphs for concept words in which the frequency of appearance of attributes shifted drastically (Figs. 4–13). In each graph, the vertical axis shows the frequency of appearance and the horizontal axis shows the timeline (the frequency of appearance is normalized so that the maximum value would be 1 and the minimum value would be 0).

Fig. 4 shows the frequency of the attribute "sports" in the concept word "Russia." The high frequency observed on December 4, 2010 is considered to have been influenced by the fact that Russia was selected as the host of 2018 FIFA World Cup on December 3, 2010. From Fig. 5, it is observed that the frequency of the attribute "economics" is high for the concept word "France" around December 19, 2010. Similar tendencies are observed for other European countries and the tendencies are considered to be the result of the influence of the EU summit. As shown in Fig. 6, which is for the concept word "Yamato," the attribute "movie" made frequent appearances in early December, 2010 influenced by the release of a film having the concept word as its title in Japan. From Figs. 7 and 8, it can be observed that the tension in the Korean peninsula resulted in frequent appearance of the attribute "international affairs" for the concept words "South Korea" and "North Korea." As shown in Fig. 9, for the concept word "fishing boat," the attribute "international affairs" recorded a high value on December 20, 2010 as a result of the ramming of a fishing boat into a South Korean Coast Guard vessel in the Yellow Sea on December 18, 2010. As shown in Fig. 10, the attribute "international affairs" showed frequent appearances on December 11, 2010 for the concept word "Nobel," influenced by the Nobel Peace Prize award ceremony held on December 10, 2010. From Figs. 11, 12, and 13, it can be observed that for the concept word "chicken," the influence of Christmas was observed in the higher-than-normal appearance of the attribute "gourmet, drink" on December 25 and 26, 2010, as well as frequent appearances of "holiday, anniversary" and "family" attributes, which are seldom observed.

From the results mentioned above, it is concluded that weblog articles are indeed influenced by various events and incidents and it is possible to extract shifts in concept attributes from tags attached to weblog articles.

IV. Conclusion

By using a concept base, which uses concept attributes that shift depending on popular topics of daily conversation, it is expected information retrieval based on the topic and construction of an associative search system will be possible. In an associative search system, more human-like association is performed. This study used weblog articles as a corpus, proposed a method for extracting concept attributes that shift as a result of various events and incidents, and exhibited the effectiveness of such a method by an experiment. Although this method is considered to be applicable to other languages because it does not rely on grammar, the results presented in this paper were obtained from an experiment performed on a weblog service based on the Japanese language. Thus, this method exhibited a strong influence from the Japanese viewpoint. For this reason, it may be necessary to select other weblog services depending on the objective, such as the creation of a concept base using international perspectives. Future issues include the study of a method to distinguish words having relatively stable concept attributes and those having unstable attributes.

Acknowledgment

We would like to thank Professors Shohachiro Nakanishi and Hiroaki Kikuchi of Tokai University for valuable discussions.

References

[1] K. Kasahara, K. Matsuzawa, T. Ishikawa, and T. Kawaoka, "Viewpoint-based measurement of semantic similarity between words," in *Proc. 5th International Workshop on Artificial Intelligence and Statistics*, 1995, pp. 56–63.

[2] K. Kasahara, K. Matsuzawa, and T. Ishikawa, "A method for judgment of semantic similarity between daily-used words by using machine readable dictionaries," *Transactions of Information Processing Society of Japan*, vol. 38, no. 7, pp. 1272–1283, 1997 (in Japanese).

[3] Y. Fujie, H. Watabe, and T. Kawaoka, "Associative document retrieval using concept-base and earth mover's distance, *IPSJ SIG Technical Reports*, 2009-ICS-154, pp. 111–116, 2009 (in Japanese).

[4] S. Tsuchiya, E. Yoshimura, and H. Watabe, "An information arrangement technique for a text classification and summarization based on a summarization frame," in *International Conf. on Natural Language Processing and Knowledge Engineering,* 2009.

[5] H. Watabe and T. Kawaoka, "Measuring degree of association between concepts for commonsense judgments," *Journal of natural language processing*, vol. 8, no. 2, pp. 39–54, 2001 (in Japanese).

[6] A. Horiguchi, S. Tsuchiya, K. Kojima, H. Watabe, and T. Kawaoka, "Constructing a sensuous judgment system based on conceptual

processing," *Lecture Notes in Computer Science*, vol. 2276, pp. 347–354, 2002.

[7] N. V. Ha, Y. Hokari, T. Ishikawa, and K. Kasahara, "A large-scale knowledge base for measuring semantic similarity between words," *IPSJ Journal*, vol. 43, no. 10, pp. 3127–3136, 2002 (in Japanese).

[8] N. Okumura, H. Watabe, and T. Kawaoka, "Extension of the concept-base applying the electronic newspaper, and the system of weight attachment to attributes," *IPSJ SIG Technical Reports*, 2005-NL-166, pp. 55–62, 2005 (in Japanese).

[9] Y. Tsuzi, H. Watabe, and T. Kawaoka, "The method of acquisition of the new concept and its attribute using the world wide Web," in *Proc. 18th Annual Conference of the Japanese Society for Artificial Intelligence*, 2D1-01, 2004, (in Japanese).

[10] K. Goto, S. Tsuchiya, H. Watabe, and T. Kawaoka, "Allocation method of an unknown search keyword to a thesaurus node by using Web," *Journal of natural language processing*, vol. 15, no. 3, pp. 91–113, 2008 (in Japanese).

[11] N. Okumura, H. Watabe, and T. Kawaoka, "A calculation method of degree of association between concepts considering fluctuation of concepts," *IPSJ SIG Technical Reports*, 2007-NL-180, pp. 79–84, 2007 (in Japanese).

[12] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[13] B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*, Addison Wesley, 2009.

[14] T. Yoshida, "A model of implicit term relationship for information retrieval," *WSEAS Trans. on Computers*, vol. 7, issue 9, pp. 1457–1466, 2008.

[15] N. Vlahovic, "Information retrieval and information extraction in Web 2.0 environment," *International Journal of Computers*, vol. 5, Issue 1, pp. 1–9, 2011.

[16] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender Systems: An Introduction*, Cambridge University Press, 2010.

[17] F. Ricci, L. Rokach, B. Shapira, P. B. Kantor, *Recommender Systems Handbook*, Springer, 2010.

[18] R. -S. Chen, Y. -S. Tsai, K. C. Yeh, D. H. Yu, and Y. Bak-Sau, "Using data mining to provide recommendation service," *WSEAS Trans. on Information Science and Applications*, vol. 5, issue 4, pp. 459–474, 2008.

[19] T. Wang, Y. Ren, "Research on personalized recommendation based on Web usage mining using collaborative filtering technique," *WSEAS Trans. on Information Science and Applications*, vol. 6, issue 1, pp. 62–72, 2009.

[20] Facebook, http:// www.facebook.com/

[21] Blogger, http://www.blogger.com/

[22] YAHOO! JAPAN Blog, http://blogs.yahoo.co.jp/

[23] YAHOO! JAPAN Developer Network, http://developer.yahoo.co.jp/

Yosuke Horiuchi was born in 1984. He received the B.E and the M.E. degrees from Tokai University, Japan, in 2008 and 2011, respectively.

In 2011, he joined Yahoo Japan Corporation. His research interests include natural language processing, social Web computing, and image processing.

Osamu Uchida was born in 1973. He received the B.E degree from Meiji University, Japan, in 1995, the M. Info. Sci. degree from Japan Advanced Institute of Science and Technology in 1997, and the Ph.D. degree from University of Electro-Communications, Japan, in 2000.

From 2000 to 2002, he was a research associate with Kanagawa Institute of Technology, Japan. He joined Tokai University, Japan, in 2002, and since 2007, he has been an associate professor with the Department of Human and Information Science, Tokai University. His research interests include information theory, image processing, Web computing, and natural language processing.

He is a member of IEEE, IEICE (The Institute of Electronics, Information and Communication Engineers), IPSJ (The Information Processing Society of Japan), IIEEJ (The Institute of Image Electronic Engineers of Japan), and JSAI (The Japanese Society for Artificial Intelligence).

Table 1: List of attributes

金融と投資 (Finance and investment)
雇用 (Employment)
ビジネス (Business)
職種 (Profession)
経済 (Economics)
インターネット (Internet)
コンピュータ (Computer)
祝日，記念日 (Holiday, anniversary)
グルメ，ドリンク (Gourmet, drink)
環境問題 (Environmental issue)
事件・事故 (Incident and accident)
災害 (Hazard)
文化活動 (Cultural activity)
季節 (Season)
映画 (Movie)
テレビ (TV)
音楽 (Music)
占い (Fortune telling)
芸能人，タレント (Actor and artist)
超常現象 (Paranormal phenomenon)
テーマパーク (Theme park)
住まい (Residence)
ペット，動物 (Pet, animal)
家庭電化製品 (Home electronics)
家庭 (Family)
政界と政治活動 (Politics and political activity)
国際情勢 (International affairs)
美容と健康 (Beauty and health)
病気，症状 (Illness)
資格試験，テスト (Qualification test)
学校 (School)
教育 (Education)
科学 (Science)
恋愛 (Romance)
結婚 (Marriage)
日本 (Japan)
世界の地方 (Region of the world)
芸術，アート (Art)
文学 (Literature)
デザイン (Design)
舞台，演劇 (Theater, drama)
人文科学 (Humanities)
スポーツ (Sports)
レジャー (Leisure activity)
趣味 (Hobby)
乗り物 (Vehicle)
ギャンブル (Gambling)

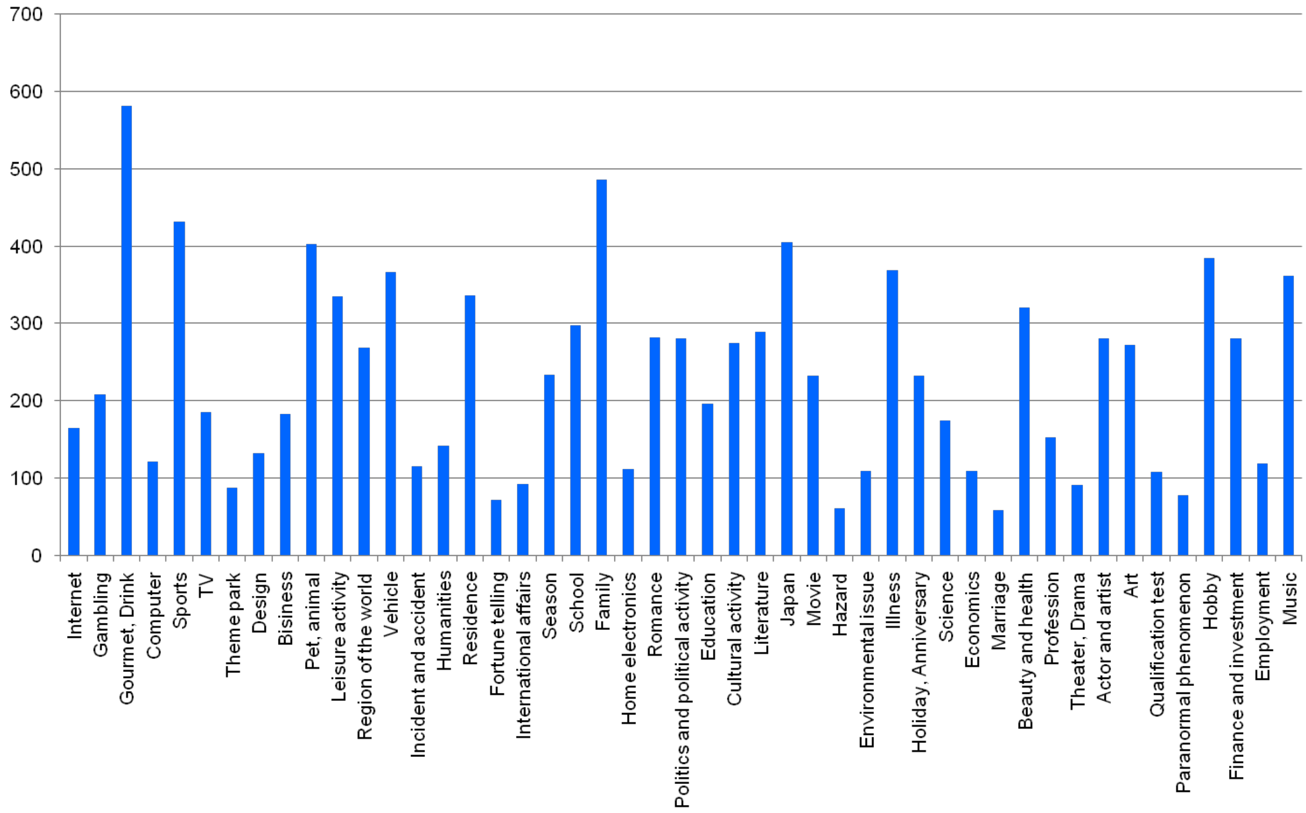


Fig. 2 Number of blog articles belonging to each category

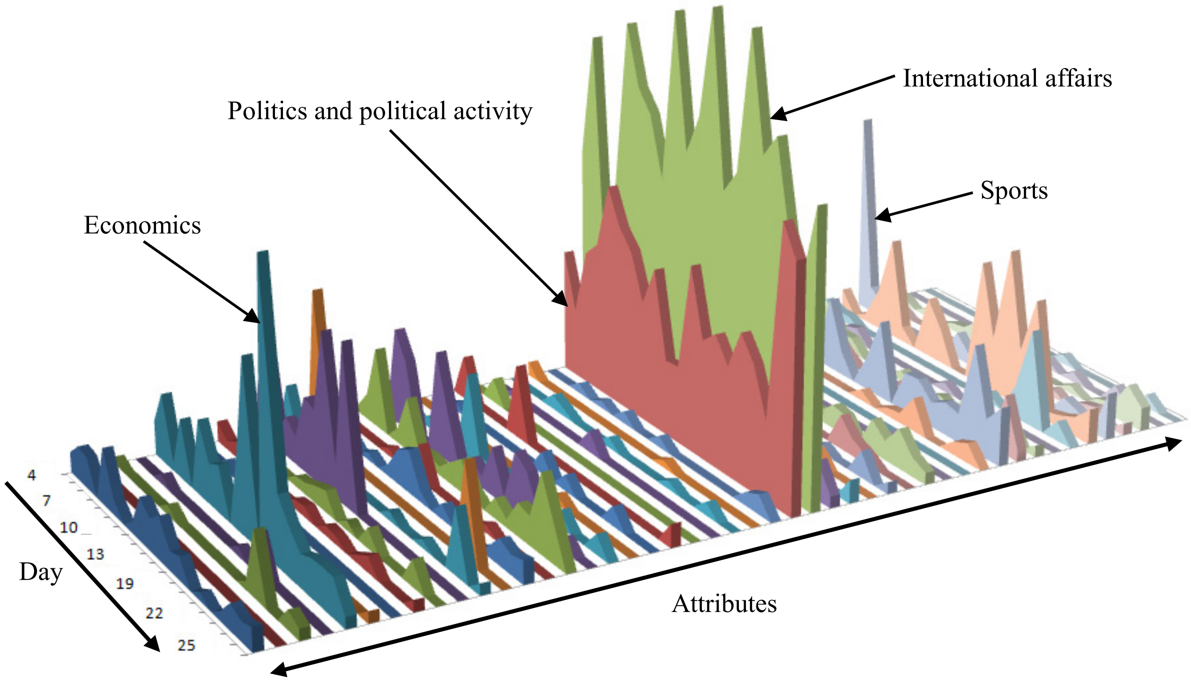


Fig. 3 Time series graph of attribute values of all attributes in the case where the concept word is “Russia”

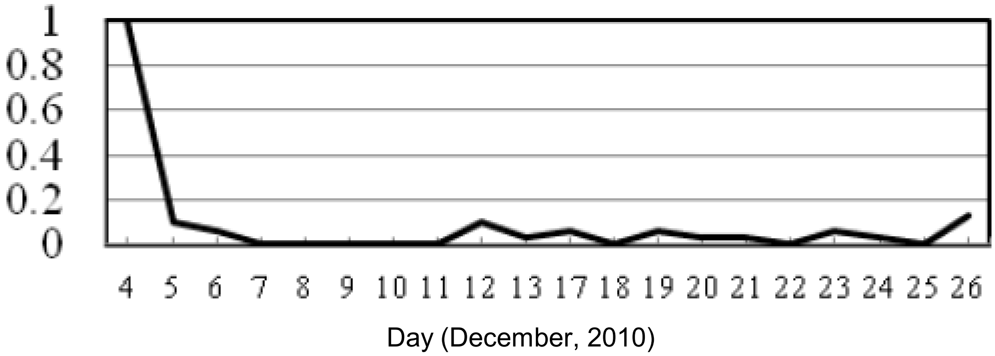


Fig. 4 Time series graph of frequency in the case where the concept word is “Russia” and the attribute is “sports”

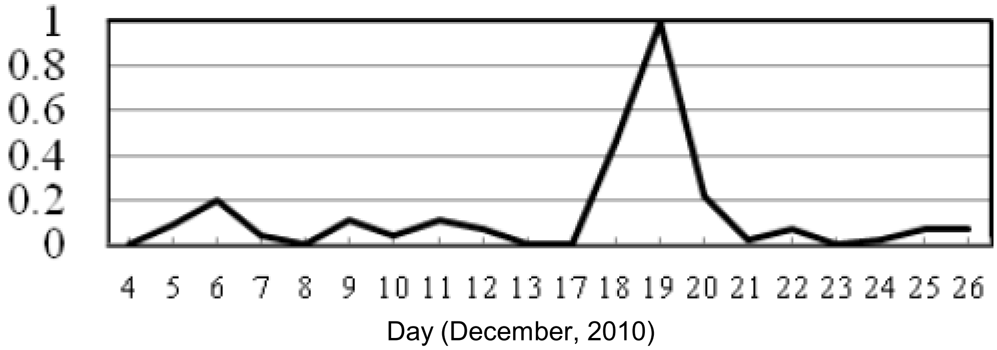


Fig. 5 Time series graph of frequency in the case where the concept word is “France” and the attribute is “economics”

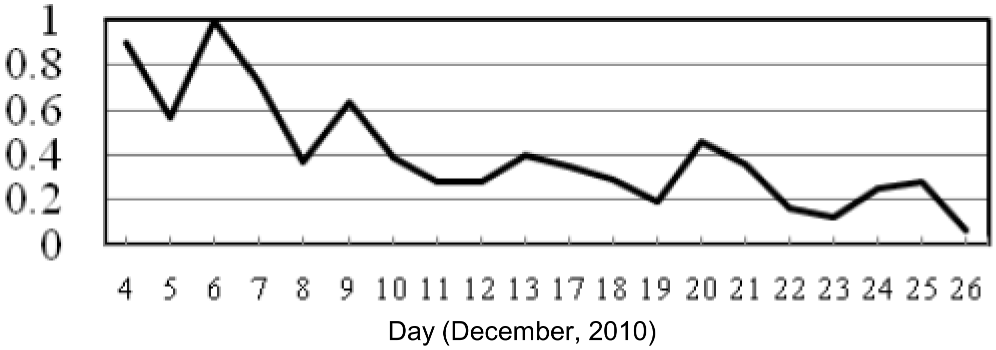


Fig. 6 Time series graph of frequency in the case where the concept word is “Yamato” and the attribute is “movie”

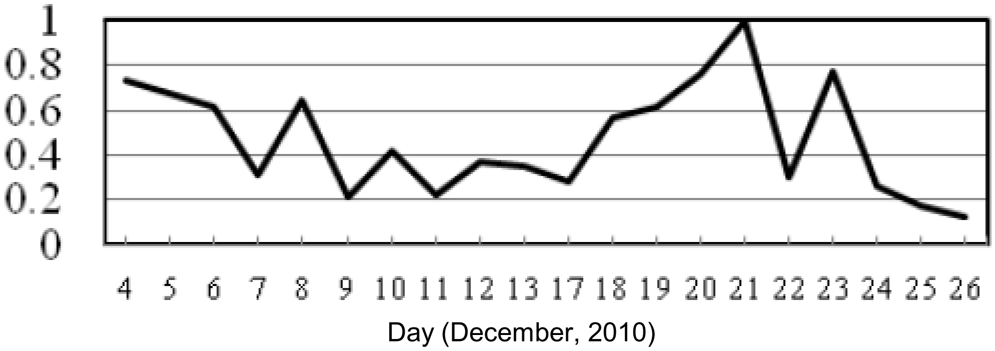


Fig. 7 Time series graph of frequency in the case where the concept word is “South Korea” and the attribute is “international affairs”

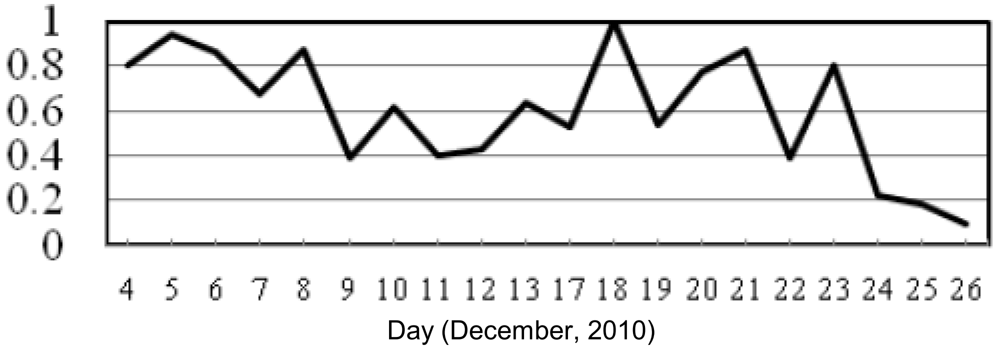


Fig. 8 Time series graph of frequency in the case where concept word “North Korea”, attribute “international affairs”

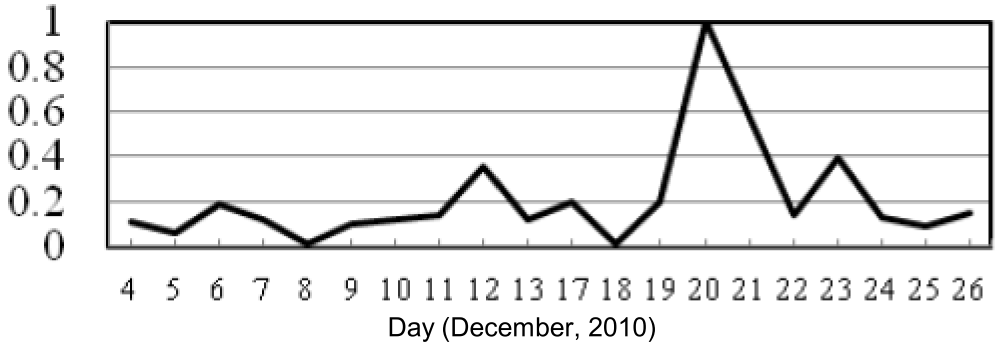


Fig. 9 Time series graph of frequency in the case where the concept word is “fishing boat” and the attribute is “international affairs”

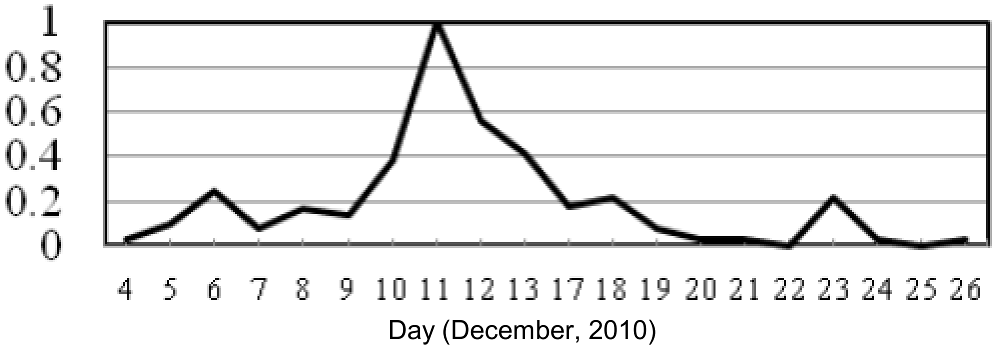


Fig. 10 Time series graph of frequency in the case where the concept word is “Nobel” and the attribute is “international affairs”

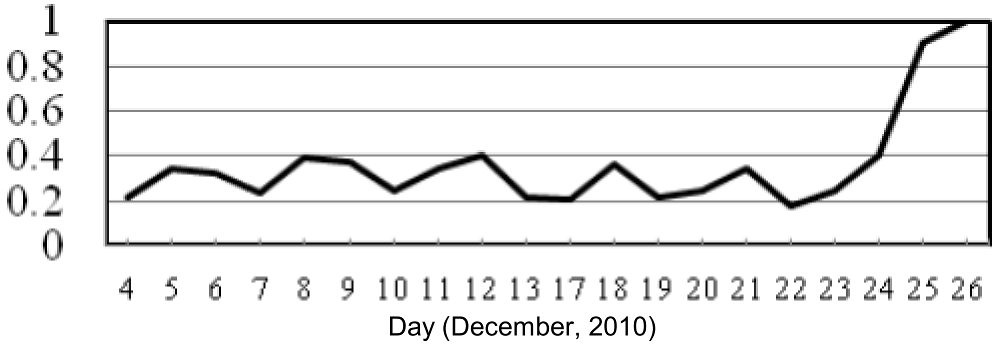


Fig. 11 Time series graph of frequency in the case where the concept word is “chicken” and the attribute is “gourmet, drink”

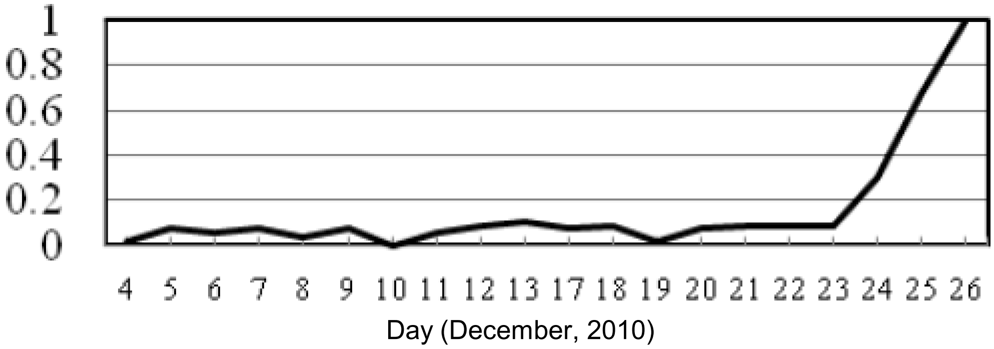


Fig. 12 Time series graph of frequency in the case where the concept word is “chicken” and the attribute is “holiday, anniversary”

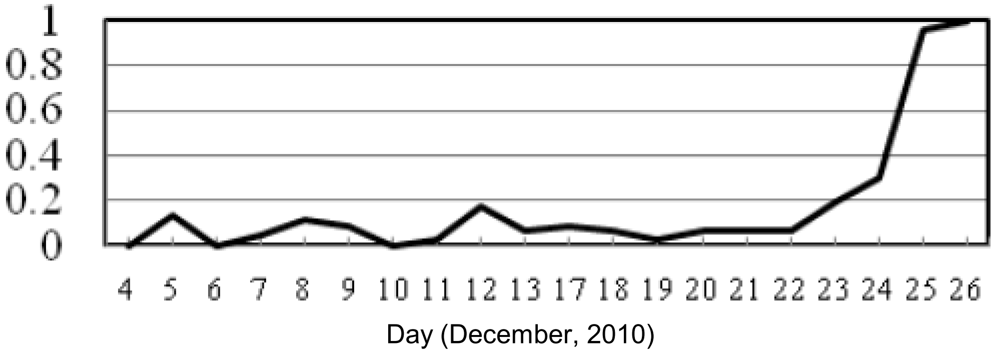


Fig. 13 Time series graph of frequency in the case where the concept word is “chicken” and the attribute is “family”