Temporal Correlation between Social Tags and Emerging Long-Term Trend Detection

Ming-Hung Hsu, Yu-Hui Chang and Hsin-Hsi Chen

National Taiwan University
R301, Dept. of CSIE, No.1, Sec.4,
Rd. Roosevelt, Taipei, Taiwan 10617
mhhsu@nlg.csie.ntu.edu.tw; hhchen@csie.ntu.edu.tw

Abstract

Social annotation has become a popular manner for web users to manage and share their information and interests. While users' interests vary with time, tag correlation also changes from users' perspectives. In this work, we explore four methods for estimating temporal correlation between social tags and detect if a long term trend emerges from the history of temporal correlation between two tags. Three types of trends are specified: steadily shifting, stabilizing, and cyclic. To compare the results of the four estimation methods, an indirect evaluation is realized by applying detected trends to tag recommendation.

Introduction

With the growth of Web 2.0, social annotation services such as del.icio.us, YouTube, and Flickr have been important manners of organizing information on the web (Hammond 2005). These Web 2.0 sites provide users with the functionality of sharing interesting and useful information with friends and even with the public, in a malleable, convenient and ease-to-use fashion. User-generated metadata in such services are often referred to as *tags*. Since tags reflect users' perception and interpretation of target resource (Li, Guo, and Zhao 2008), here we consider tags as interesting concepts for users.

The rapid popularization of social tagging has attracted considerable works for analysis or utilization of such rich metadata (Bao et al. 2007; Golder and Huberman 2006; Halpin et al. 2007; Wu et al. 2006;). In these applications, estimation of semantic correlation between two tags (or, concepts) is fundamental and indispensable. While users' interest may shift as time goes on, correlations between various concepts may shift from users' perspectives. The following example clarifies this idea. It concerns a hot topic in del.icio.us "website design programming". Intuitively users may be interested in different programming languages in various periods. As a result, the semantic correlation between "website design" and a specific programming language (e.g., "PHP") vary with time.

As far as our knowledge, the time factor is not concerned in previous researches on social tagging. In this work, we attempt to estimate temporal correlation between two concepts and explore if there is any long-term trend emerging from the history of considered temporal correlation. An emerging trend indicates how user interest varies with time. There are three main issues to be addressed for this novel problem: 1) How to model the history of temporal correlation between two tags in an efficient manner? 2) What types of long-term trends are reasonable to be detected? 3) How to detect the emerging trend of a specified type?

The fundamental assumption of this work is: *appearance of an annotation reflects the condition that at the time when the annotation was generated, the annotator was interested in the concepts she assigned*. Temporal correlation between two concepts is thus estimated by using the co-occurrence information. We model the evolutional history of concept correlation over a long period by estimating temporal correlation between two concepts in each sliding time frame in the period. We then specified three types of emerging trends in the history of temporal correlation: *steadily shifting*, *stabilizing*, and *cyclic*. For the task of trend detection, we employed typical regression models with various predictor functions.

Data Preparation

Our dataset of annotations was crawled from del.icio.us during February 2008. Only the partial set generated before 2007/12/31 is used for trend detection, and annotations generated in January 2008 was used for evaluation. The data for trend detection include annotation histories for 337,187 distinct URLs. These annotation records are generated by 1,197,440 unique users between January 2004 and December 2007, 48 months throughout. There are 91,600,046 annotations totally. In the dataset, the time is

divided into months and thus discrete. On average, each annotation record includes 4.36 tags.

The annotation data is very sparse particularly when the concerned time interval in which the annotations are generated is short. To reduce the harmfulness of data sparseness and to filter out the noises introduced by various tag strings of identical concept, we performed preprocessing on tag strings as follows.

- *Case transformation.* Tags are transformed to lower case.
- *Removing punctuation marks.* All punctuations in a tag string are removed.
- *Stemming.* We perform Porter's stemming algorithm to identify terms in different morphological forms.
- *Removing rare concepts.* As our goal is to discover the implicit long-term trends, we remove the rare concepts by *popularity* and *occurrence frequency*. A concept is said to have enough popularity if it is used to annotate at least 5 URLs in a month. On the other hand, a concept has enough frequency of occurrence if it has enough popularity in 12 or more months between January 2004 and December 2007.

After preprocessing, 21,335 concepts remain for the detection task.

Temporal Correlation and Emerging Trend

Temporal Correlation Estimation

In this study, length of a time frame is fixed as three months. We compare four methods for estimating temporal correlation between two concepts. These methods are proposed based on various *degrees of strictness*. The following are basic notations for the four methods:

- $corr_M(c_i, c_j, t)$ = the correlation between concepts c_i and c_j, estimated by a specific method M in the time frame t
- $RS(c_i, t)$ = the set of URLs assigned with concept c_i included in annotations generated in t
- $CS(r_j, t)$ = the set of concepts, included in annotations generated in t, assigned to URL r_j
- $fr(c_i, r_j, t)$ = the frequency of URL r_j assigned with concept c_i in t

The first method follows set overlap, named **OVL**:

$$corr_{\mathrm{OVL}}(c_i, c_j, t) = \frac{RS(c_i, t) \cap RS(c_j, t)}{RS(c_i, t) \cup RS(c_j, t)} \quad (1)$$

The other three methods are based on typical cosine similarity in vector space models, except that the weighting schemes are varied, i.e.

$$corr_M(c_i, c_j\ t) = \frac{\vec{V}_{c_i,t} \cdot \vec{V}_{c_j,t}}{\left|\vec{V}_{c_i,t}\right| \cdot \left|\vec{V}_{c_j,t}\right|} \quad (2)$$

where $\vec{V}_{C_i}$ represents the vector of concept c_i. whose the jth element is $w(c_i, r_j)$. To determine $w(c_i, r_j)$ in Formula (2), the weight of URL r_j in $\vec{V}_{C_i}$, the following schemes are proposed. The name of each weighting scheme denotes the corresponding method M.

CF: $w(c_i, r_j)$ is the frequency of c_i assigned to r_j, i.e.

$$w(c_i, r_j, t) = fr(c_i, r_j, t) \quad (3)$$

NCF: $w(c_i, r_j)$ is $fr(c_i, r_j)$ normalized by the sum of $fr(c_k, r_j)$ for $c_k \in CS(r_j)$, i.e.

$$w(c_i, r_j, t) = \frac{fr(c_i, r_j, t)}{\sum_{c_k \in CS(r_j, t)} fr(c_k, r_j, t)} \quad (4)$$

The idea of NCF is to treat all resources fairly.

RNCF: a reweighted version of NCF. A URL which received more concept assignments would obtain a heavier weight than other URLs in the vector, i.e.

$$w(c_i, r_j, t) = \frac{fr(c_i, r_j, t)}{\sum_{c_k \in CS(r_j, t)} fr(c_k, r_j, t)} \cdot \log \sum_{c_k \in CS(r_j, t)} fr(c_k, r_j, t) \quad (5)$$

Detecting Emerging Trends

As a preliminary study, we specify three types of intuitive long-term trends that may emerge from the evolution history of temporal correlations: *steadily shifting* (*SS*), *stabilizing*, and *cyclic*. Figure 2 illustrates two examples for each type respectively. For each trend type, we employ a linear regression model with specified predictor functions to detect whether the concerned evolution of temporal correlation is involved (Neter, Wasserman, and Kutner 1990). That is, suppose Y denotes the modeled trend line and X denotes the time, then

$$Y = f(X) = \sum_i \lambda_i P_i(X) \quad (6)$$

where $P_i(X)$ denotes the predictor function and λ_i is the combination coefficient in the regression model of *least square error*.

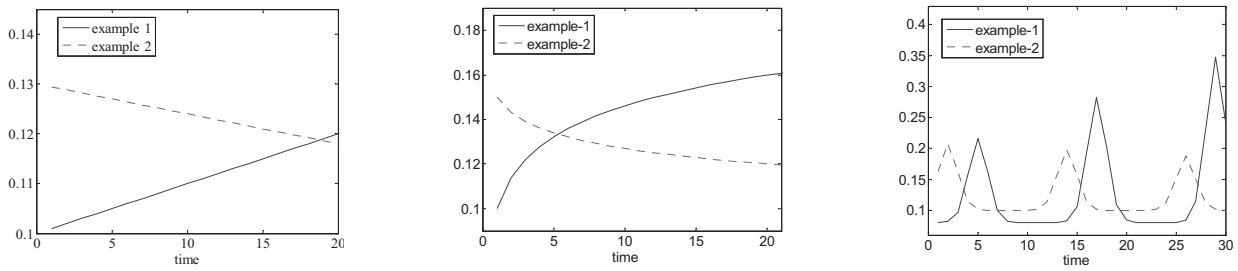
The predictor functions $P_i(X)$ are described as follows:

- Steadily-shifting trends: $P_1(X) = 1$, and $P_2(X) = X$
- Stabilizing trends: $P_1(X) = 1$, and $P_2(X) = \log(X)$
- Cyclic trends: $P_1(X) = 1$, $P_2(X) = e^{\beta \cos(\frac{2\pi}{L}(X\ \ \theta))}$, and $P_3(X) = X \cdot e^{\beta \cos(\frac{2\pi}{L}(X\ \ \theta))}$.

As Figure 2(c) shows, a cyclic trend implies periodic peaks in the history of temporal correlations. L was set to 12 to detect annual trends and β was empirically set to 4. According to the fitness defined by the R-square coefficient (Draper and Smith 1998), a modeled trend line is detected if its fitness is higher than the threshold.

Experiment on Evaluating Emerging Trends

As the detected trends cannot be evaluated directly, we perform an indirect evaluation by integrating detected



(a) (b) (c)

Figure 1. Illustration of the three types of trends: (a) steadily shifting; (b) stabilizing; (c) cyclic.

trends into an application, tag recommendation, to compare the four proposed methods for estimating temporal correlation. Suppose the set of tags annotated on URL r_k currently is S_k. For each candidate tag c_i in the training data, its recommendation score based on its similarities with tags in S_k is defined as follows:

$$score(c_i) = \prod_{c_j \in S_k} sim(c_i, c_j) \quad (7)$$

where $sim(c_i, c_j)$ can be computed by typical methods such as cosine similarity or point-wise mutual information. We use the detected emerging trends to modulate $sim(c_i, c_j)$.. The modulation function is:

$$AF(x) \quad \left(\frac{f(x)}{\bar{y}} \right)^{\alpha} \quad (8)$$

where $\bar{y}$ is the average of concerned temporal correlations and α is introduced as a parameter to control how sensitive this modulation is. $f(x)$ is the detected trend.

The testing data for tag recommendation include 4,658 URLs in our dataset. Each testing URL received at least 50 annotations during January 2008. For each URL, tags assigned by at least two del.icio.us users in January 2008 compose the set of relevant tags. For each testing URL, the last one annotations received in December 2007 were used as the basic information S_k. Recommendation results are evaluated by precision at top 10 (P@10) and precision-recall break-even point (BEP).

Experiment Results and Discussions

Table 1 shows the number of emerging trends for each method, when the fitness threshold was set to 0.7. OVL and NCF propose the most and the least numbers of candidate trends, respectively. Recall that NCF has the highest degree of strictness and OVL has the lowest. The total number of proposed emerging trends is related to the strictness of the temporal correlation estimation method. Even when OVL was adopted, the total number of emerging trends, 804,210, is much fewer than the number of possible concept pairs (21,335*21,334/2) from 21,335 concepts, indicating that for most concept pairs (or user interest from our viewpoint), no long-term trends are detected.

Method	Trend Type			Total
	SS	Stabilizing	Cyclic	
OVL	236,599	561,885	5,726	804,210
CF	7,084	128,462	5,700	141,246
NCF	7,304	18,972	15,658	41,934
RNCF	33,362	56,932	24,003	114,297

Table 1. Number of detected emerging trends

As it is impracticable to directly evaluate which trends proposed by which methods reflect real variations of user interests, Table 2 shows the effects of utilizing detected trends to modulate concept similarities for the tag recommendation task. The baseline (BL) method is typical cosine similarity. We set α to various values, obtained similar results, and only show performances of two settings here due to the space limit. Compared with the baseline, the performance is improved only when integrating trends emerging from RNCF. The improvement is significant with 99% confidence in paired two-tails *t*-test. Among the results of 4,658 testing URLs, 540 results are improved, 390 results are damaged and the others are invariable, when α was set to 0.5. On the other hand, trends emerging from OVL show great damage on recommendation performance, which indicates that OVL estimated temporal correlation loosely and imprecisely, resulting in that many unrealistic trends emerged. Moreover, the comparison between results of integrating with NCF and with RNCF implies that URLs which received more annotations in a time frame were more indicative to reflect user interest and provided more information for temporal correlation estimation from users' perspective.

Method	α 1		α 0.5	
	BEP	P@10	BEP	P@10
BL	.4017	.4619	.4017	.4619
BL+OVL	.0876	.1025	.1107	.1948
BL+CF	.4016	.4620	.4017	.4621
BL+NCF	.3998	.4587	.4010	.4608
BL+RNCF	**.4031**	**.4634**	**.4040**	**.4636**

Table 2. Performance of tag recommendation integrating detected emerging trends

Figure 2 shows the recommendation performance in BEP when integrating with trends emerging from RNCF, versus the fitness threshold of trend detection. When the threshold dropped from 0.95 to 0.7, the number of positive emerging trends increased and larger improvement in recommendation was obtained. When the threshold dropped to be lower than 0.7, the performance dropped due to that many unrealistic and noisy trends were introduced.

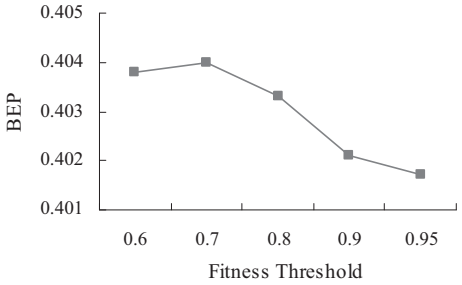


Figure 2. Influence of the fitness threshold for trend detection on recommendation performance in BEP

Figure 3 shows an example of detected cyclic trend emerging from RNCF, which involves the concept pair of "gift" and "egg". The fitness is 0.7356 and the peak occurs in April annually. This trend is supposed to be greatly associated with Easter. Some other cyclic trends are discovered involved concept pairs associated with Christmas, with the peaks occurring in December. For example pairs of "gift" and "wrapper", "xmas" and "e-shopping", "xmas" and "advice", "xmas" and "home-decoration" are with such trends. However, we observe that most detected cyclic trends are hard to be recognized or explained, indicating that we may need more constraints to filter out noises when detecting cyclic trends.

Table 3 shows some concepts whose temporal correlations with "google" or "facebook" are detected as *arising* trend of steadily-shifting type. To consider a concept pair as a topic, an arising trend detected in the history of temporal correlation indicates that delicious users are more and more interested in the concerned topic.

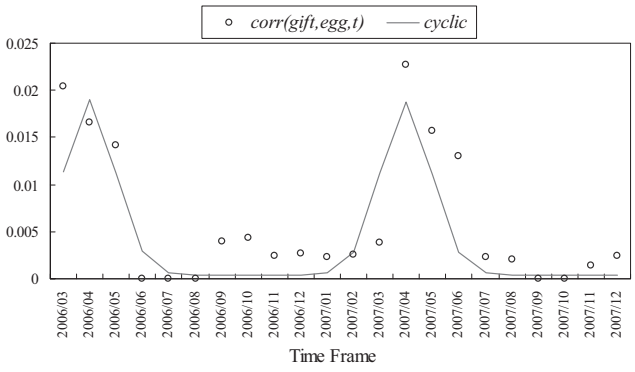


Figure 3. Temporal correlation between "gift" and "egg", estimated by RNCF

Conclusion and Future Work

In this study, we explore a novel issue of detecting emerging long-term trend in temporal correlation between two social tags. We propose four methods: OVL, CF, NCF, and RNCF for estimating temporal correlation and adopt regression models with various predictor functions to detect the three types of emerging trends: steadily-shifting, stabilizing and cyclic. The results are indirectly evaluated by integrating the detected trends to adjust tag similarities for tag recommendation. Significant improvement in recommendation performance is obtained when RNCF is adopted. For future work, we are investigating more constraints to detect cyclic trends more precisely.

	Associated Concepts
google	advertising, analytics, api, bookmark, calendar, code, css, delicious, doc, download, editor, freeware, gadget, hadoop, jquery, lecture, lifehack, mac, outlook, product, rss, …
facebook	analysis, bbc, blogpost, bot, contact, develop, digitalmedia, flash, interact, lifestream, music, network, opensource, outlook, photo, php, ruby, share, socialnetwork, todo, web2, video, …

Table 3. Concepts associated with "google" or "facebook" with arising trends of steadily shifting type

Acknowledgement

This work is supported partially by National Science Council under the contract NSC96-2628-E-002-240-MY3.

References

Bao, S. et al. 2007. Optimizing Web Search Using Social Annotations. *Proceedings of the 16th International Conference on World Wide Web*.

Draper, N.R. and Smith, H. 1998. *Applied Regression Analysis*. Wiley-Interscience, ISBN 0-471-17082-8

Golder, S.A. and Huberman, B.A. 2006. Usage Patterns of Collaborative Systems. *Journal of Information Science*. 32(2), 198-208.

Halpin, H. et al. 2007. The Complex Dynamics of Collaborative Tagging. *Proceedings of the 16th International Conference on World Wide Web*.

Hammond, T., Hannay, T., Lund, B. and Scott, J. 2005. Social Bookmarking Tools (I): a General Review. *D Lib Magazine.* 11(4).

Li, X., Guo, L., and Zhao, Y. 2008. Tag-based Social Interest Discovery, *Proceedings of the 17th International Conference on World Wide Web*.

Neter, J., Wasserman, W., Kutner, M.H. 1990. *Applied linear statistical models : regression, analysis of variance, and experimental designs (3rd edition)*. IRWIN, Illinois. ISBN 0-256-08338-X

Wu, X. et al. 2006. Exploring Social Annotations for the Semantic Web. *Proceedings of the 15th International Conference on World Wide Web*.