Analysis of User's Relation and Reading Activity in Weblogs

Tadanobu Furukawa[1], Tomofumi Matsuzawa[2], Yutaka Matsuo[3],
Koki Uchiyama[4] and Masayuki Takeda[2]

[1] Graduate School of Science and Technology, Tokyo University of Science,
2641 Yamazaki, Noda-shi, Chiba, Japan
[2] Dept. of Information Sciences, Tokyo University of Science,
2641 Yamazaki, Noda-shi, Chiba, Japan
[3] National Institute of Advanced Industrial Science and Technology,
2-41-6 Aomi, Koto-ku, Tokyo, Japan
[4] hottolink, Inc.
2-11-17 Nishigotanda, Shinagawa-ku, Tokyo, Japan

Abstract. In a blog network, there are many relations such as comment, trackback, and so on. We consider that if the relations are related to user's reading activity, we can extract useful information from the relations for using a recommendation system. We define the strength and type as the measure for relations, and analyze the correlation between those measures and users' reading activity. We attempt to determine the relations on which users regularly read the blogs.

1 Introduction

Much information describing individuals exists on the WWW as diaries, news sites and Bulletin Board Systems (BBSs) because of the popularization of internet services. Therefore, it is difficult for users to get exactly the information they seek, so "information retrieval" or "information recommendation" services are in demand.

Recently, Weblogs (blogs) are receiving attention as a new form of a personal information transmission. Blog characteristics include: users update their contents frequently because the operations to information sending are easy to do on a web browser – basically one blog is updated by one user; it has functions called "comment" and "trackback"; and it offers the updated information in the form of an RDF Site Summary (RSS)[7]. Particularly trackback, a unique function to blog, allows interactive communication and allows the use of discussion of one chosen topic.

Blog users visit others' blogs and write comments or send trackbacks as they update their own blogs. Considering these activities, we can discover the relationships among blogs. Actually, such relations are closely watched in business. Particularly, blogs are often read with social networking services (SNSs), which are services that record and map human relations information in recent years.

We notice the relations among blogs and analyze their effects on users' activities. First, we verify the hypothesis "whether users who visit blogs are strongly related". We prepare various relations such as Comment / Trackback, and inspect those relations that appear to be influential. Next, we clarify "whether we can distinguish blogs that users read frequently using the relations". If possible, we might build a recommendation service based on relations among blogs.

Our analyses use the database: "Doblog"[5] , a blog-hosting service in Japan. Using this service, users update their blogs and perform other activities such as writing comments after they log in. Therefore, we can treat users' behavior. Doblog has a special function called "bookmark", to link to favorite blogs from one's own blog. We use it as one relation. We analyze users' activities using this data for limited users on that service.

The subsequent section explores related works. Section 3 explains the algorithms used for analysis. in Section 4, we show a concrete experiment technique. Section 5 describes analysis of the results. We conclude the paper in Section 6.

2 Related Works

Characteristics of hyperlinks on WWW were investigated in detail[1], and research of extracting useful information as webpages or web communities of a specific topic from the network structure based on

[5] ©NTT Data Corp., ©hottolink,Inc.,
http://www.doblog.com/ , using data of Oct. 2003 – Jul. 2004

hyperlinks was successful[5][6]. For example, it is known that the WWW community can be extracted as a complete bipartite subgraph comprising a hub and authority.

However, the blog network is very complex because of its unique relations such as Trackback and Comment. For that reason, we cannot easily grasp the state of information spread[3].

This paper presents analysis of this complex network and our attempt to extract useful information from it. Particularly, one strong point is that we use users' behavior as one relation.

3 Relations between two Blogs

We think there must be the blogs a user can access hardly by following hyperlinks but those are interesting to the user. We try to distinguish such blogs by using the relations with own blog, prepare some relations: Bookmark / Comment / Trackback. If user A, the author of blog A, makes Bookmark link to blog B in own blog, we define there is relation of Bookmark from A to B. If user A send Comment / Trackback to blog B more than once, we define there is relation of Comment / Trackback from A to B.

Herein, if user A has visited user B's blog more than once, we call it "Visiting (relation/activity)" from A to B; if user A has visited user B's blog more than 30 times, we call it "Regular Reading (relation/activity)" from A to B. The logs of user's "Visiting" are available from the Doblog database. We analyze whether correlation exists among the relations of the two blogs and Visiting / Regular Reading.

Our goal is to extract those blogs that a user wants to visit repeatedly after one access based on the relations that a user cannot access easily. Therefore, we check the two points below: (i) What relations engender Visiting behavior? (ii) What relations on Visiting relation engender Regular Reading?

Regarding the relations of blogs, we use (a) link structure using Bookmark / Comment / Trackback, (b) similarity of fields of interest.

For (a), we check the range of two hops because it seems that the most influential relation to users' activity is a two-hop relation in the relations we want to extract, i.e. indirect relations. We analyze this two-hop relation for "Strength" and "Type".

– Strength of Relation
  A measure of each relational strength is represented by the number of routes that connect two blogs in two hops by one relation. For example, routes between A and B are three ($A-1-B/A-2-B/A-4-B$), as shown in Fig. 1. Figure 2 shows the extent to which the rate follows the transitivity rule ($A-1 \cap 1-B \Rightarrow A-B$) in these two hops. This figure reveals a positive correlation between the number of routes and the transitivity rule. For that reason, we can say that this number indicates the strength of the relations. We examine the correlation between this strength measure and Visiting / Regular Reading activity.

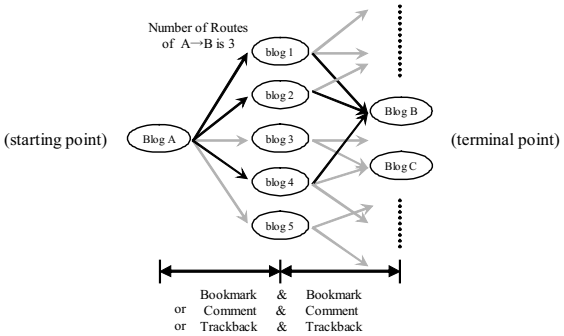

Fig. 1. Number of routes.

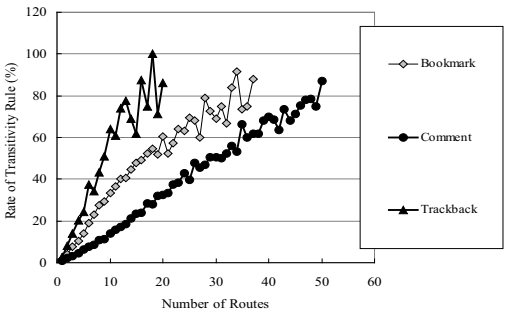

Fig. 2. Number of routes & transitivity rule.

– Type of Relation
  As Fig. 3 shows, 12 kinds of relation exist in the range of two-hop relations (gray nodes in the figure). We analyze in which relation Visiting / Regular Reading occurs readily for this 12 kinds .

With regard to (b), this uses the items[6] that users selected at the Doblog user registration. As a

[6] 103 items such as sports, entertainment, etc.

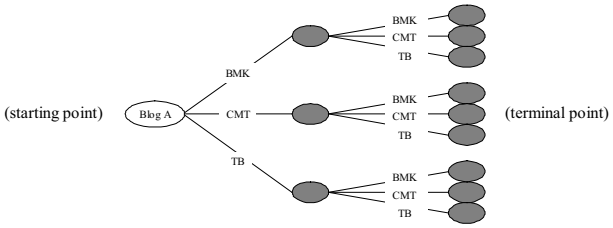

Fig. 3. 12 kinds of relations. (BMK, Bookmark; CMT, Comment; TB, Trackback)

measure of strength among users, we use the number of identical items. We quantify the correlation between strength and Visiting / Regular Reading activity.

We analyze the above using the UserRank top 1647 blogs, 10% of all Doblog users. UserRank is calculated from monthly data of Bookmark, Comment, Trackback, and Visiting; it is a score based on how the user contributes to the network constructed. The nodes are blogs (users) and edges are relations using the spreading activation[2]. Therefore, it seems that using highly ranked blogs, i.e., blogs which have many relations to other blogs, allows more effective experiments. These are heavy users of blogs: only 1647 users' data among all data occupy 59% of Bookmark data, 64% of Comment data, and 64% of Trackback data.

4 Experiment Algorithm

We conducted the following experiment to examine the Visiting / Regular Reading (we call those two activities "Reading" in this chapter) in each relation under the conditions below; we also calculate the rate. The Reading Rate is calculated using the following formula:

$$Reading\ Rate = \frac{Number\ of\ Blog\ Sets\ which\ have\ Reading\ Relation}{Number\ of\ All\ Blog\ Sets}$$

"Number of all blog sets", in the Visiting analysis, indicates the number of all blog sets of the starting and terminal points under each condition. In the Regular Reading analysis, it represents the number of the blog sets under each condition and the Visiting relation.

4.1 Regarding Strength of Relation

Number of Routes. For the two-hop relation, which consists of one Bookmark / Comment / Trackback, we calculate the Reading Rate for each number of routes.

1. One hop: Extract all blogs connected from one blog (the starting point) by one relation.
2. Two hops: Extract all blogs (the terminal point) connected from blogs in operation 1 using the same relation as operation 1, except for the starting point.
3. Check the number of routes and whether there is a Reading relation between the starting point and each terminal point.
4. For all starting points, repeat operations 1 to 3.
5. Calculate the Reading Rate according to the number of routes.

Similarity of Interests. For all two-blog sets, check the number of fields of same interests and whether they have a Reading relation. This relation does not have direction. However, considering users A and B, we found the Reading relation for $A - B$ and $B - A$. We checked about ${}_{1647}P_2 = 2,710,962$ sets, and calculated the Reading Rate for each number of the same fields of interest.

4.2 about Type of Relation

For all two-blog sets (${}_{1647}P_2$ sets), we checked which relation among the 12 kinds of relations mentioned in the previous chapter existed, and whether they have a Reading relation (Table 1). Using each set as training data, we analyzed which relation determines the Reading activity, and constructed a decision tree using a machine learning algorithm C4.5[4]. During Visiting analysis, we analyzed the above only for the sets showing a two-hop relation because there must be a Visiting relation between one-hop relation sets.

Table 1. Form of training data for analysis of relation types.

2 blogs		12 Types of Relation					Visit
start	term	Bmk	Cmt	...	Tb-Cmt	Tb-Tb	(Read)
A	B	T	F	...	F	T	TRUE
A	C	F	T	...	F	F	FALSE
B	C	T	T	...	T	T	TRUE
⋮				⋮			⋮

Table 2. Form of training data for analysis of relation strength.

2 blogs		Number of Routes			Visit
start	term	Bmk	Cmt	Tb	(Read)
A	B	3	3	1	TRUE
A	C	5	6	3	TRUE
B	C	4	2	0	FALSE
⋮			⋮		⋮

(start/term, the starting/terminal point; Bmk, Bookmark; Cmt, Comment; Tb, Trackback; Read, Regular Reading relation; Visit, Visiting relation; T, True; F, False)

5 Results and Analysis

This chapter explains results of the experiments explained in the previous chapter.

5.1 Strength of Relation

for Visiting Activity. The relation between the number of routes and identical fields of interest and Visiting activity is shown in Fig. 4. This figure shows that

- If the number of the routes increases, the Visiting rate rises.
- There is no correlation between the number of fields of identical interest and Visiting activity.

Regarding the former, the reason seems to be that the Bookmark / Comment / Trackback creates a hyperlink that allows a user to visit the site easily.

Regarding the latter – similarity of interest and Visiting –, we inferred the following:

- Users do not always record their interests in their own blogs. Alternatively, they do not always visit blogs that contain information that is related to their selected fields of interest.
- If not above, users have not been able to follow the blogs of interest.
- The users do not select the fields of interest carefully.

This will be argued more in the following section.

Regular Reading Activity. Figures 5 is the diagram showing the relation between strength of relation and Regular Reading activity. This figure shows that

- if the number of the routes increases, the Regular Reading rate rises.
- There is no correlation between the number of identical fields of interest and the Regular Reading rate.

The latter is supportive of the second guess of the same interest-Visiting analysis in the previous section. However, validating the users' written contents about their interests is the work of text mining, so this study ignores that analysis.

Decision Tree. Using Figs. 4 – 5, Trackback seems to be the most influential relation for Visiting / Regular Reading activity, but we cannot determine a hierarchy among relations, because there may be more than two kinds of relations between blogs. Therefore, using these analyzed data about the strength of relation, we construct decision trees for Visiting / Regular Reading. For Visiting analysis, the training data are 568,043 sets of two blogs(starting point and terminal point) that have more than one route by any two-hop relation; for Regular Reading analysis, they are 154,549 sets of two blogs that have a Visiting relation. Each data set for machine learning consists of the number of routes of each relation and whether a Visiting / Regular Reading relation exists (Table 2). These data do not include the number of identical fields of interest because that seems to show no correlation.

Results are shown in Figs. 6 – 7. Figure 6 depicts the analyses of Visiting. Figure 7 is analysis of Regular Reading.

In these decision trees, the highly influential relation occupies the upper position as a node, and B / C / T represents Bookmark / Comment / Trackback. The left branch is a path for which there

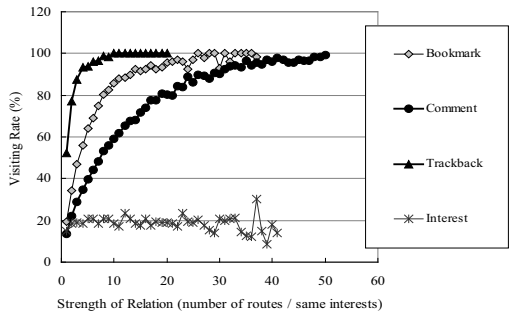

Fig. 4. Number of routes or similarity of interest & Visiting.

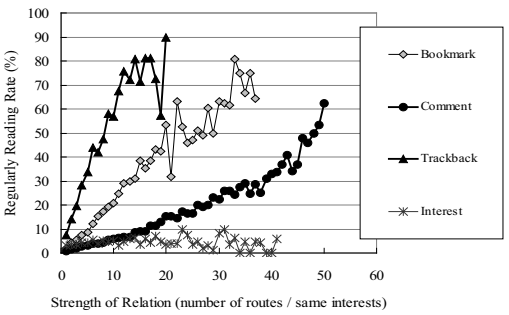

Fig. 5. Number of link routes or similarity of interest & Regular Reading (more than 30 visits).

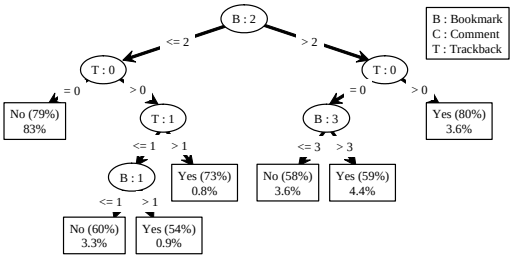

Fig. 6. Decision tree of Visiting by number of routes. (568,043 training sets).

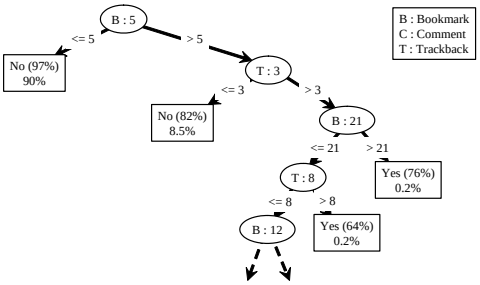

Fig. 7. Decision tree of Regular Reading by number of routes. (154,549 training sets).

are less than the number of label routes. The right branch is a path fo which there are more than the number of label routes. The leaves appear when it is classified to some extent. "Yes" and "No" represent whether the Visiting / Regular Reading relation is true or false; the values represent (right) the rate of correct distinction and (below) the percentage of classified data compared to the whole. For example, following the left part of the diagram down in Fig. 6, the rate of blog sets that have less than two routes of Bookmark and less than zero route(= no routes) of Trackback is 83%. They do not have Visiting relations with a probability if 79%. It is noteworthy that these show only the parts of the top five layers because of the space limitations of this paper.

Figure 7 shows that the most influential relation to Regular Reading activity is the number of the routes of Bookmark. Trackback follows and it is influential, inspite of its smaller number of routes. However, Trackback relation is not visible (if blog A send Trackback to blog B, there is a hyperlink of B to A, but there is not a hyperlink of A to B), so it is not clear why the two-hop Trackback routes is influential. The rate of Bookmark between two blogs which have Trackback is not very high, $2826/5192 = 54.4\%$.

On the other hand, Fig. 6 shows that the most influential relation to Visiting activity is also the number of routes of Bookmark, and the number of routes is very small(more than two). This says that users usually visit blogs by following Bookmarks. And this number is smaller than the one of Regular Reading analysis, so the probability of to visit blogs on the relations which are influential to Regular Reading activity. Because the more numerous the routes, the higher the Visiting probability according to Fig. 4. So, we cannot get relations worth recommending in this analysis.

Table 3. The average number of each relation routes.

Relation	Bookmark	Comment	Trackback
average	1.20	1.20	0.14

5.2 Type of Relation

for Visiting Activity. Result showing which two-hop relation is influential on Visiting activity is Fig. 8.

The decision tree's structure is almost identical to that of the previous section. Two letter relations such as "CT" represent a two-hop relation: Comment – Trackback. The branches represent whether there is a relation (left is false, right is true).

The most influential relation to Visiting activity is the Trackback – Trackback relation, this is a similar result of Strength of Relation.

for Regular Reading Activity. Result shows the two-hop relations that are influential for Regular Reading activity is in Fig. 9.

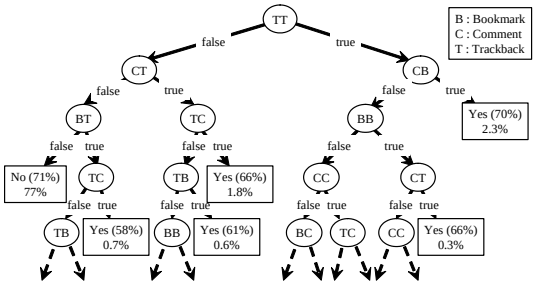

Fig. 8. Decision tree of Visiting by 19 kinds of relations. (323,861 training sets).

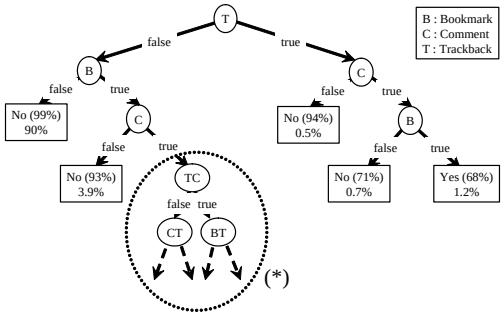

Fig. 9. Decision tree of Regular Reading by 19 kinds of relations. (206,804 training sets).

Table 4. The rate of each two-hop relation existence.

Relation	BB	BC	BT	CB	CC	CT	TB	TC	TT
rate(%)	26.3	31.2	13.2	27.6	29.9	14.2	13.7	21.1	6.2

The figure shows that the most influential relation for Regular Reading activity is the one-hop relation, such as Bookmark / Comment / Trackback. The sets of two blogs that have neither a Trackback nor a Bookmark relation occupy about 90% of all data and have almost no Regular Reading relation. This fact shows the large effect of one-hop relation (Users generally send Trackback / Comment or make Bookmark link to the blogs he regularly reads).

As the effect of one-hop relation is too big, we pay attention to only two-hop relation (the part of (*) in Fig. 9): the relations of (1) "Trackback - Comment" is true and "Bookmark - Trackback" is true and (2) "Trackback - Comment" is false and "Comment - Trackback" is true seem to be influential to Regular Reading. Now, we culculate the Visiting rate of two relations, they are (1)64% and (2)44%. So these relations may be worth recommending.

6 Conclusions

In this study, we analyzed a blog network: which blog users visit and regularly read, by exploring its unique relations. We defined the strength and type of the relations as measures and tried to elucidate the influence of those relations on the users' reading activities.

Consequently, for the Visiting activity: the two-hop Bookmark route is influential, and wheter the two-hop Trackback route exists is also effective. For Regular Reading activity: in analysis of Strength we could not get useful result. However, in analysis of Type we could find the relation, that is influential for Regular Reading but user cannot surely visit. We consider that such analyses can form the basis of information recommendation.

References

1. A.Broder, R.Kumar, F.Maghoul, P.Raghavan, S.Rajagopalan, R.Stata, A.Tomkins, and J.Wiener: Graph Structure in The Web. The 9th International. World Wide Web Conference (2000)
2. A.Saad: A Multi-Agent Spreading Activation Network Model for Online Learning Objects. http://julita.usask.ca/mable/saad.pdf (2001)
3. E.Adar, L.Zhang, L.A.Adamic, and R.M.Lukose: Implicit Structure and the Dynamics of Blogspace. http://www.hpl.hp.com/research/idl/papers/blogs/blogspace-draft.pdf (2003)
4. J.R.Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, Inc. (1993)
5. L.Page, S.Brin, R.Motwani, and T.Winograd: The PageRank Citation Ranking:Bringing Order to the Web. http://web.resource.org/rss/1.0/spec (1999)
6. R.Kumar, P.Raghavan, S.Rajagopalan, and A.Tomkins: Trawling the Web for Emerging Cyber-Communities. The 8th International World Wide Web Conference (2001)
7. RSS-DEV Working Group: RDF Site Summary (RSS) 1.0. http://web.resource.org/rss/1.0/spec (1999)