A Link-Based Rank of Postings in Newsgroup

Hongbo Liu[1], Jiahai Yang[1], Jiaxin Wang[2], and Yu Zhang[2]

[1] The Network Research Center
[2] Department of Computer Science and Technology
Tsinghua University, Beijing, China, 100084
liuhb1@gmail.com

Abstract. Discussion systems such as Usenet, BBS, Forum are important resources for information sharing, view exchanging, problem solving and product feedback, etc. on Internet. The postings in newsgroups on Usenet represents the judgments and choices of participators. The structure of postings could provide helpful information for the users. In this paper, we present a method called PostRank to rank the postings based on the structure of newsgroup. Its results correspond to the eigenvectors of the transition probability matrix and the stationary vectors of the Markov chains. It could provide useful global information for the newsgroup and it can be used to help the users access information in it more effectively and efficiently. This method can be also applied on other discussion systems. Some experimental results and discussions on real data sets collected by us are also provided.

Keywords: link analysis, rank, newsgroup, discussion systems.

1 Introduction

Usenet is a world-wide distributed discussion system, and it is one of the representative information resources on Internet. Usenet provides a convenient way for the communication and organization of discussions, which is much different from the World Wide Web (WWW) whose main purpose is information publishing. It consists of a set of newsgroups with names classified hierarchically by subject. In each group, postings are posted to the NNTP server and broadcast to other servers. With these servers, people all over the world can subscribe newsgroups they are interested in and can participate in the discussions.

Comparing with web pages, the content of postings on Usenet is generally more informal, brief and personalized. It contains rich information and ideas contributed by the participators. Due to the huge size of Usenet, people can only subscribe a few groups and generally read a small fraction of the postings. It may take quite much time for a newbie to familiar with a group and use it sensibly. It is difficult to access the required information efficiently from all the postings in a group because of its huge size and loose organization.

Information Retrieval (IR) techniques have been used on the web to make information more accessible. IR techniques have widely used the "bag-of-words model" for tasks such as document matching, ranking, and clustering [1]. On

P. Perner (Ed.): MLDM 2007, LNAI 4571, pp. 392–403, 2007.

WWW, because of the intrinsic hyperlink property of web pages, link analysis based on ideas of social networks has also been used in the ranking system of some search engines [2,3]. The simplicity, robustness and effectiveness of link-based ranking method have been witnessed with the success of Google, whose basis of ranking system is PageRank. Social networks have also been applied in other domains, such as marketing [4], email relationship [5], chat [6] and so on [7].

As most postings on the Usenet do not contain hypertexts, they can not be benefited from these link-based algorithms of WWW directly. Some IR techniques were used to improve the services of Usenet [8]. The briefness and casualty of newsgroup postings make it difficult for conventional text mining techniques. Some investigations based on social networks have also been done to extract useful information from the Usenet [9,10].

On Usenet, it is not easy to choose the postings before read when the users browse the newsgroup. Normally it is needed to read the postings throughout thread to get related information. It is time consuming and many postings are not very valuable. Some hints of postings may be greatly helpful to improve the efficiency of the Usenet users. For the Usenet search, the order of results is very important, which may be improve with a good ranking system. Therefore, good posting rank with intrinsic properties of a newsgroup can make the information on Usenet more accessible.

In this paper, according to the characteristics of Usenet, a link-based method to calculate the rank of postings on Usenet is proposed. Some mathematical analysis of this method is discussed and experimental results on real data sets are also given.

2 The Calculation of PostRank

2.1 Usenet Newsgroup and Its Representation

Unlike web pages, the postings on a newsgroup of Usenet are organized by threads. Each thread is invoked by one seed posting and followed by several response postings. The quantity and content of postings are determined by collaborative work of the participators along with the evolution of discussions.

Considering the posting v_i as node and the respondent relationship $e = \langle v_{i1}, v_{i2} \rangle$ as link, each thread can be abstracted as a rooted tree whose root is the seed posting and its descendants are the response postings of this thread. In the rooted tree, the leaf nodes are the postings without response. Since there are many threads in a newsgroup, its structure can be represented with forest $G(V, E)$. In this way, a newsgroup contains m postings with s seed postings and t leaf postings could be represented as a forest with m rooted trees and $m - s$ links.

Supposing posting v_i have c_i neighbors, due to the tree structure of thread there are two classes of neighbors for v_i according their relationships with v_i, i.e., the parent set $\mathcal{P}(v_i)$ containing a_i postings, and the offspring set $\mathcal{O}(v_i)$

containing b_i postings. a_i may be 0 or 1 depending on whether v_i is a seed posting. From the above description, we can get that

$$a_i + b_i = c_i, \sum_{v_i \in G} c_i = 2(m - s), \sum_{v_i \in G} |a_i - b_i| = 2t. \tag{1}$$

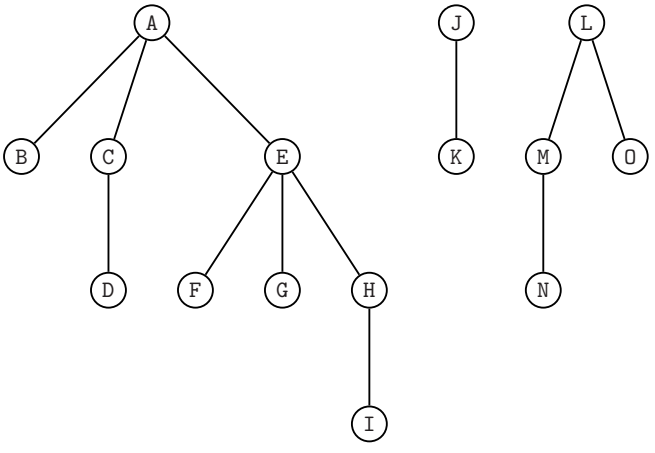

Fig. 1. graph representing a small newsgroup containing 3 threads with 15 postings

A small newsgroup containing 15 postings, including 3 seed postings and 8 leaf postings was shown in Fig. 1. It can be represented by the following adjacent posting matrix

$$\mathbf{M} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \\ I \\ J \\ K \\ L \\ M \\ N \\ O \end{array} \begin{array}{c} \begin{array}{cccccccccccccccc} A & B & C & D & E & F & G & H & I & J & K & L & M & N & O \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

According to the above descriptions, the following properties can be inferred directly.

Property 1. *(1) The diagonal elements of* $\mathbf{M}$ *are zero. (2)* $\mathbf{M}_{ij}$ *satisfies*

$$\sum_{j=1}^{m} \mathbf{M}_{ij} = a_i, \sum_{i=1}^{m} \mathbf{M}_{ij} = b_j, \sum_{i=1}^{m} \sum_{j=1}^{m} \mathbf{M}_{ij} = m - s$$

.

In this manuscript, the rooted trees are denoted with $Tk, k = 1, 2, \cdots, s$, and we use the symbols with subscript Tk to denote parameters of tree Tk. For tree Tk, posting is represented as $v_{i|Tk}$. The number of postings in Tk is m_{Tk}, and the number of leaf postings in Tk is t_{Tk}, and the level of Tk is l_{Tk}.

2.2 PostRank Calculation and Analysis

One intuitive idea to rank the postings based on G is that if a posting was responded by more postings, its rank is higher. Thus, b_i could be a candidate of the rank of v_i. This seems to be simple and feasible. However, this calculation works with the assumption that all postings have equal contribution to the rank, which is not reasonable enough. All postings are not created equal. A posting responded by a high rank posting might be more valuable than posting with a low rank response posting. It is better to retrieve the rank from the Usenet structure recursively.

Similar with some link-based rank methods on the web, the link from v_i to v_j in the forest G can be viewed as vote from v_i to v_j. If the rank of a posting is high, the postings it responding to and responded by provide some information intimately related with it, which might also be valuable for the users. When the users browse the Usenet, it is natural to read the parent or the offspring of his interested posting to get more detailed information. Thus, in our calculation, posting $v_{i|Tk}$ votes other postings according to their relationships with $v_{i|Tk}$. In the reverse direction, the rank of $v_{i|Tk}$ is determined by the ranks and relationships with other postings based on the structure of Usenet. Since the seed postings and the leaf postings are special in the Usenet, we add self-loop to them to give them additional bonuses. The rank of posting $v_{i|Tk}$ may contribute to its parent, its offspring, other postings in the same thread, and any posting in the newsgroup differently. In PostRank, we use $\alpha, \beta, \lambda, \eta$ to describe the difference of these relationships. Therefore, the PostRank $r_{i|Tk}$ of $v_{i|Tk}$ can be calculated as following:

$$r_{i|Tk} = \alpha \sum_{v_j \in \mathcal{P}(v_{i|Tk})} r_j / b_j + \beta \sum_{v_j \in \mathcal{O}(v_{i|Tk})} r_j + \lambda \sum_{v_j \in Tk} r_j / m_{Tk} + \eta, \qquad (2)$$

where $0 \leq \alpha, \beta, \lambda, \eta < 1$ and $\alpha + \beta + \lambda + \eta = 1$.

As the PostRank vector $\boldsymbol{r}^T$ is a m dimensional row vector, it is convenient to use the matrix form of Eq. (2) when calculating. To illustrate this easily, two transformations were introduced. In the following discussions, by $\mathscr{M}, \mathscr{V}$ we denote matrix space and vector space respectively.

Transformation $\mathcal{D} : \mathscr{M} \to \mathscr{M}$ is defined as follows to distill the self-loops of seeding postings or leaf postings according to the posting matrix $\mathbf{M}$. For $\mathbf{A} \in \mathscr{M}$,

$$\mathcal{D}_{ij}(\mathbf{A}) = \begin{cases} 0 & \text{if } i \neq j \\ 0 & \text{if } i = j \text{ and } \sum_{j=1}^{m} \mathbf{A}_{ij} \neq 0 \\ 1 & \text{if } i = j \text{ and } \sum_{j=1}^{m} \mathbf{A}_{ij} = 0. \end{cases}$$

According to the definitions and Eq. (1), $\mathcal{D}(\mathbf{M})$ is the adjacent matrix of root self-loops and the $\mathcal{D}(\mathbf{M}^T)$ the adjacent matrix of leaf self-loops. Let square matrix

$$\mathbf{T} = \mathbf{M} + \mathcal{D}(\mathbf{M}). \tag{3}$$

$\mathbf{T}$ represents the new graph of G plus self-loops of root nodes. Since there is only one parent or self-loop for each node, each row of $\mathbf{T}$ contains and only contains one nonzero element.

Let l_{max} be the max level in G, i.e., $l_{max} = \max(l_{Tk})$. The whole thread Tk was included in the columns of seed postings of square matrix $\mathbf{T}^{l_{max}}$, and $\mathbf{T}^{l_{max}}$ could be used to indicate the correspondent relationships of postings and their seed postings.

Transformation $\mathcal{N} : \mathscr{M} \to \mathscr{M}$ is defined as the normalization of the matrix row vectors based on their l_1 norms. That is, for $\mathbf{A} \in \mathscr{M}$,

$$\mathcal{N}_{ij}(\mathbf{A}) = \mathbf{A}_{ij} / \sum_{j=1}^{n} \mathbf{A}_{ij}.$$

Therefore, the matrix form of PostRank can be represented with

$$\boldsymbol{r}^T = \boldsymbol{r}^T(\alpha\mathbf{T} + \beta\mathcal{N}(\mathbf{M}^T + \mathcal{D}(\mathbf{M}^T)) + \lambda\mathcal{N}(\mathbf{T}^{l_{max}})) + \eta\boldsymbol{w}^T, \tag{4}$$

where $\boldsymbol{w}^T$ is a m dimensional personalized row vector with $\boldsymbol{w} > 0$. $\boldsymbol{w}$ could be used to customize the PostRank vector for special demand. In Eq. (2), $\boldsymbol{w}^T = \boldsymbol{e}^T$, where $\boldsymbol{e}$ be a m dimensional column vector with all ones.

For the implement of PostRank calculation, Eq. (4) can be written in the form of iteration as

$$\boldsymbol{r}^T(k) = \boldsymbol{r}^T(k-1)(\alpha\mathbf{T} + \beta\mathcal{N}(\mathbf{M}^T + \mathcal{D}(\mathbf{M}^T)) + \lambda\mathcal{N}(\mathbf{T}^{l_{max}})) + \eta\boldsymbol{w}^T. \tag{5}$$

For the implementation of Eq. (5), its convergent property should be considered.

Property 2. *Let* $\mathbf{P} = \alpha\mathbf{T} + \beta\mathcal{N}(\mathbf{M}^T + \mathcal{D}(\mathbf{M}^T)) + \lambda\mathcal{N}(\mathbf{T}^{l_{max}})$. $\alpha + \beta + \lambda$ *is the spectrum radius of* $\mathbf{P}$.

Proof. Based on the above descriptions and Property 1,

$$\begin{aligned} \sum_{j=1}^{m} \mathbf{P}_{ij} &= \alpha \sum_{j=1}^{m} \mathbf{T}_{ij} + \beta \sum_{j=1}^{m} \mathcal{N}(\mathbf{M}_{ij}^T + \mathcal{D}(\mathbf{M}_{ij}^T)) + \lambda \sum_{j=1}^{m} \mathcal{N}(\mathbf{T}_{ij}^{l_{max}})) \\ &= \alpha + \beta + \lambda, \end{aligned} \tag{6}$$

so we have $\mathbf{P}\boldsymbol{e} = (\alpha + \beta + \lambda)\boldsymbol{e}$. Therefore $\alpha + \beta + \lambda$ is the eigenvalue of $\mathbf{P}$ and $\boldsymbol{e}$ is the corresponding right eigenvector.

According to the matrix property, the spectrum radius of $\mathbf{M}$

$$\begin{aligned}\rho(\mathbf{P}) &\le \|\mathbf{P}\|_\infty = \max_{i,j} \mathbf{P}_{ij} \\ &\le \alpha \max_{i,j}(\mathbf{T}_{ij}) + \beta \max_{i,j} \mathcal{N}(\mathbf{M}_{ij}^T + \mathcal{D}(\mathbf{M}_{ij}^T)) + \lambda \max_{i,j} \mathcal{N}(\mathbf{T}_{ij}^{l_{max}}) \qquad (7) \\ &= \alpha + \beta + \lambda.\end{aligned}$$

Considering Eq. (6) and Eq. (7), $\rho(\mathbf{P}) = \alpha + \beta + \lambda$.

Starting from non-zero initial vectors $\boldsymbol{r}^T(0)$, $\boldsymbol{r}^T(k)$ can be calculated based on $\boldsymbol{r}^T(k-1)$ using Eq. (5). Because $\rho(\mathbf{P}) = \alpha + \beta + \lambda < 1$ from Property 2, PostRank calculation can converge to their stable vector $\boldsymbol{r}^T$, which is the solution satisfying Eq. (4).

The meaning of PostRank can also be understood with the discrete Markov model. Defining square matrix

$$\mathbf{Q} = \mathbf{P} + \eta \boldsymbol{e}\boldsymbol{w}^T, \qquad (8)$$

when $\|\boldsymbol{r}^T(0)\|_1 = 1$, Eq. (5) can be written as

$$\boldsymbol{r}^T(k) = \boldsymbol{r}^T(k-1)\mathbf{Q}. \qquad (9)$$

When $\|\boldsymbol{w}^T\|_1 = 1$, according to Property 2 we have

$$\mathbf{Q}\boldsymbol{e} = \mathbf{P}\boldsymbol{e} + \eta \boldsymbol{e}\boldsymbol{w}^T\boldsymbol{e} = (\alpha + \beta + \lambda)\boldsymbol{e} + \eta\boldsymbol{e} = \boldsymbol{e}. \qquad (10)$$

Thus, $\mathbf{Q}$ is a stochastic matrix and the PostRank calculation build a Markov chain with transition probability matrix $\mathbf{Q}$. Since $\eta > 0$, from Eq. (8) $\mathbf{Q}$ is primitive. Hence the Markov chain can converge to its stationary vector, that is, the PostRank vector $\boldsymbol{r}^T$. The Markov chain indicates random walk model that as the Usenet user read one posting, he may jump to the posting it responded to, the posting it responded by, any posting in the same thread or any posting in the newsgroup with different probability on the next. When $\|\boldsymbol{w}^T\|_1 = 1$, with this model PostRank vector is the stationary probability distribution of all postings. From Eq. (10), PostRank vector is also the eigenvector corresponding to eigenvalue 1 of $\mathbf{Q}$ which was constructed on the newsgroup structure using PostRank equation. Therefore, PostRank vector $\boldsymbol{r}^T$ could reflect the nature features of G, and it is the intrinsic property and good measure of postings in a newsgroup.

In $m \times m$ matrix $\mathbf{M}$, there are only $m - s$ nonzero elements, which makes $\mathbf{M}$ very sparse. $\mathbf{P}$ can be obtained based on $\mathbf{M}$ before iteration using the definition of Property 2 and $nnz(\mathbf{P}) \le 3m$, where $nnz(\mathbf{P})$ is the number of non-zeros in $\mathbf{P}$. The process of Eq. (5) iteration is matrix-free and only $nnz(\mathbf{P})$ multiplications are needed for each step. Only the storage of one vector $\boldsymbol{r}^T(k)$ is required at each iteration. Thus, this algorithm is suitable for the large size and sparsity of the posting matrix of Usenet newsgroup. Some experiments were performed to acquire the PostRank vector on realistic datasets. The experimental results achieved will be discussed in the next section.

```
From: diffuser78@gmail.com
Newsgroups: comp.lang.python
Subject: Re: OS specific command in Python
Date: 21 Jun 2006 06:34:42 -0700
Organization: http://groups.google.com
Lines: 40
Message-ID: <1150896882.746722.95200@u72g2000cwu.googlegroups.com>
References: <1150781429.090359.148560@c74g2000cwc.googlegroups.com>
   <1150783324.258644.65770@u72g2000cwu.googlegroups.com>
   <4498dcd5$0$25503$626a54ce@news.free.fr>
NNTP-Posting-Host: 66.255.187.74
Mime-Version: 1.0
Content-Type: text/plain; charset="iso-8859-1"
X-Trace: posting.google.com 1150896887 8314 127.0.0.1 (21 Jun 2006 13:34:47 GMT)
X-Complaints-To: groups-abuse@google.com
NNTP-Posting-Date: Wed, 21 Jun 2006 13:34:47 +0000 (UTC)
In-Reply-To: <4498dcd5$0$25503$626a54ce@news.free.fr>
User-Agent: G2/0.2
Xref: news.edisontel.com comp.lang.python:41439
```

Fig. 2. The header of a typical posting on Usenet

3 Experiments and Their Results

3.1 Datasets Preparation

We wrote a bot program in Perl to download the postings from the NNTP server. The bot program communicates with NNTP server using socket connection following RFC 977 specification [11] and save the postings in text file with Mailbox format. Since only the headers are needed in our calculation, the headers were separated from the postings from the Mailbox file, and they are stored using CSV format after some text treatment. The header of a typical posting is shown in Fig. 2. The contents in CSV file were ordered and imported to the database. SQL statements were performed on the database by a Java program through JDBC interface to construct and extract the structure of newsgroup based on the header information. Postings in uncompleted thread were removed from the data sets during the structure extraction. The process of data sets collection and newsgroup structure extraction was shown in the diagram of Fig. 3.

Experiments were preformed on two data sets collected from comp.lang.perl.misc and comp.lang.python, which are two active newsgroups about computer languages on Usenet. The datasets are called DS1 and DS2 in the following.

DS1 contains 10532 postings including 1286 participators and 1774 threads of comp.lang.perl.misc from Mar 5, 2006 to Jun 27, 2006. DS2 contains 18821 postings including 2463 participators and 3408 threads of comp.lang.python from Mar 5, 2006 to Jun 27, 2006.

The probability distributions of response posting number b are shown in a log-log plot of Fig. 4. In DS1 and DS2, a few postings got many response postings and a lot of postings were only responded by few response posting or not responded. From the figure, the distributions exhibit power-law feature of $P(b) = b^{-\gamma}$ with $\gamma \simeq 4.1$ for both DS1 and DS2. Power-law distribution has also discovered on Internet and other systems[12], and it means the heterogeneity of network structure which is helpful for our PostRank.

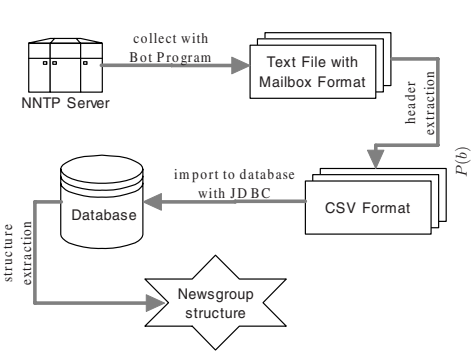

Fig. 3. The process of data collection and structure extraction of Usenet newsgroup

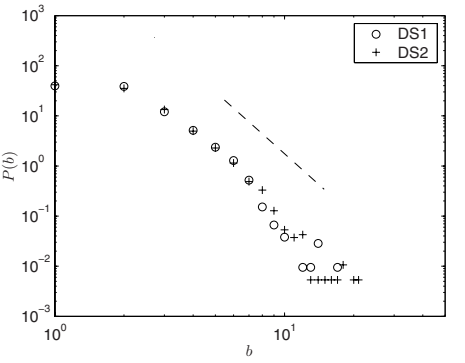

Fig. 4. The probability distribution of response posting number of postings in DS1 and DS2, the slope of dashed line is −4.1

3.2 Experimental Results

In our experiments, PostRank vector $\boldsymbol{r}^T$ was calculated with $\alpha = 0.25, \beta = 0.45, \lambda = 0.15, \eta = 0.15$. These parameters are determined from our experiments, and they can be adjusted to change the impacts of different kinds of postings. The personalized vector $\boldsymbol{w}^T$ was assigned $\boldsymbol{e}^T$, so according to Eq. (4) the l_1 norm of PostRank vector $\|\boldsymbol{r}^T\|_1 = \|\boldsymbol{w}^T\|_1 = \|\boldsymbol{e}^t\|_1 = m$.

We measure the rates of convergence using the l_1 norm of the residual vector, i. e.,

$$\Delta^{(k)} = \|\boldsymbol{r}^T(k) - \boldsymbol{r}^T(k-1)\|_1.$$

The convergence rates of in our experiments of DS1 and DS2 were plotted on a semi-log graph shown in Fig. 5. It could converge rapidly, which follows $O((\alpha + \beta + \lambda)^k)$.

Since most PostRank scores are small, the logarithms of PostRank of DS1 and DS2 are shown in the histograms of Fig. 6. In these histograms, there are few postings with high PostRank, and many PostRank scores are around the average value 1. Comparing with Fig. 4, we can see that they are very unalike. Effected by the number and PostRank scores of different kinds of related postings simultaneously, r_i is quite different with b_i for posting v_i. The relationship r_i and b_i is shown in Fig. 7, where each symbol represents a posting and the cycle symbols and plus symbols denote postings from DS1 and DS2 respectively. In this Figure, we could see that postings with same b_i may be very different in r_i, and vice versa. In DS1, the highest PostRank 37.145 is obtained by a seed posting titled "What is Expressiveness in a Computer Language", which was responded by only 4 postings with high PostRank scores. The thread it invoked contains 467 response postings, but this seeding posting is not very significant barely considering b_i. According to this figure, we get results alike for DS2.

In our example above, some seed postings with few b_i rank high mainly because they have a lot of descendants. The number of direct and indirect

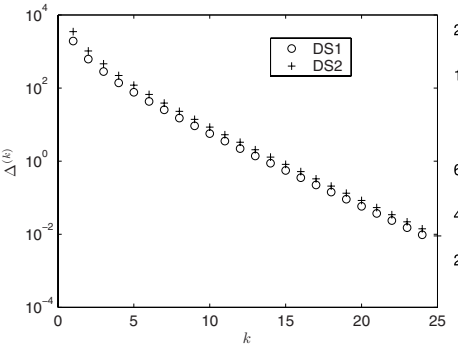

Fig. 5. The convergence rates of PostRank calculation of DS1 and DS2

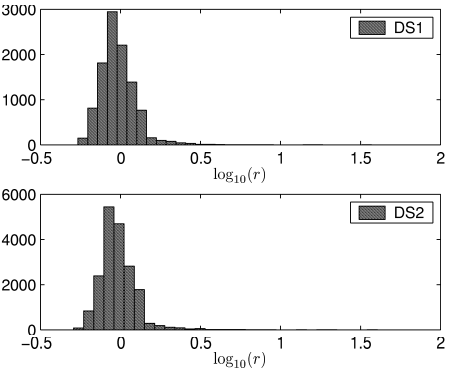

Fig. 6. The PostRank histograms of postings in DS1 and DS2

descendants can be used as the candidate for the rank, which could also makes the seed postings rank high. However, this will lead to unreasonable results that parents always rank higher than their descendants. In PostRank, there is the opportunity for the descendants rank higher than their parents, such as when the offspring of descendants have high ranks. The influence of parents, descendants and other postings were considered simultaneously according to their distances and relationships with the posting being ranked, so PostRank is found to be a good ranking method for the postings in newsgroup.

As we discussed, PostRank can provide useful clues based on the newsgroup structure for the users to help them access the information more effectively. It can also be used in other applications of Usenet data mining. For example, we can obtain some helpful properties of the participators based on PostRank. On the Usenet, participators are judged only by his postings, irrespective of his social status or appearance. They are the soul of a newsgroup. The participators behave very differently owing to their character and knowledge background. Acquaintance and evaluation of participators in a newsgroup are very important for the users to use the newsgroup effectively. However, it may take quite much time, so some hints of participators may be of great help for this.

In the newsgroup, suppose there are n participators represented as $p_u, u = 1, 2, ..., n$. All postings posted by p_u is $\mathcal{T}(p_u)$, and the number of postings in $\mathcal{T}(p_u)$ is d_u. Define f_u^{sum} as the sum of PostRank of postings in $\mathcal{T}(p_u)$ and f_u^{ave} as the average PostRank of $\mathcal{T}(p_u)$, i.e.,

$$f_u^{sum} = \sum_{v_i \in \mathcal{T}(p_u)} r_i, f_u^{ave} = f_u^{sum}/d_u.$$

The relationship of f_u^{ave} and f_u^{sum} was shown in the log-log plot of Fig. 8 where each symbol represents a participator. The results from DS1 and DS2 are plotted in subgraphs respectively. In this figure, for the participators with high f_u^{sum} the average values of f_u^{ave} are about 1. Many participators with high f_u^{ave}

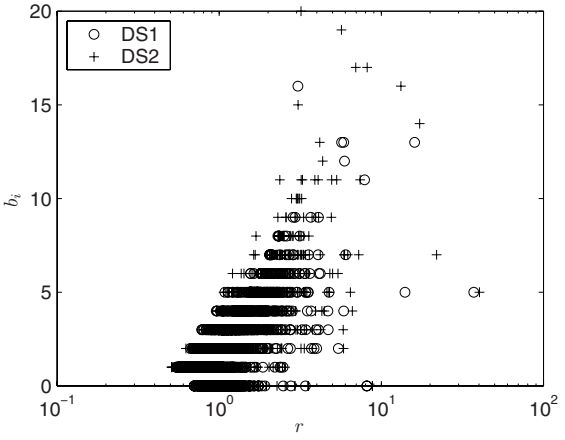

Fig. 7. Comparison of number of response posting and PostRank for postings in DS1 and DS2

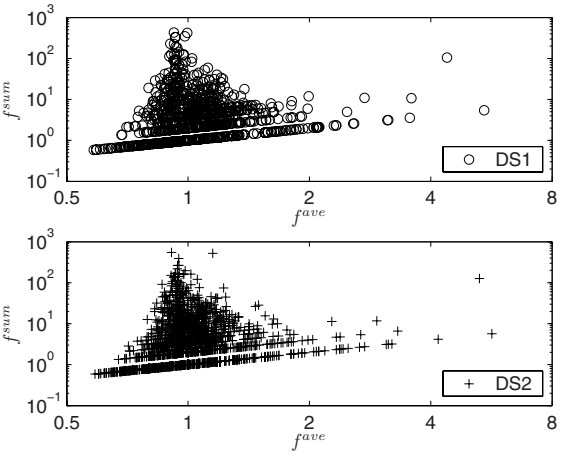

Fig. 8. The relationship of f^{sum} and f^{ave} of participators in DS1 and DS2

have small f_u^{sum}, which means they posted only a few postings. On the upper right of DS1 plot in Fig. 8, there is one participator whose f_u^{sum} and f_u^{ave} are both high, which make him relatively special. He is xah@xahlee.org, who own a homepage http://xahlee.org on computer and scientific art which was created in 1996 and visited by about 7 to 10 thousand unique visitors per day. Among 24 postings he posted in DS1, 4 postings get PostRank more than 10. Barely considering d_u, he is easy to be neglected. From Fig. 8, similar results can also be obtained for DS2.

4 Conclusions

In this paper, we proposed a method to calculate the PostRank vector of postings in a newsgroup based on the newsgroup structure. From the analysis, we can see that our method can converge rapidly. Its results correspond to the eigenvectors of the transition probability matrix and the stationary vectors of the Markov chains.

The calculation of PostRank is link-based and content independent, and it can be computed offline using only the posting headers. Therefore, it can be implemented on the servers of newsgroup services or on the newsgroup client softwares. It could provide useful intrinsic attribution for the postings and can be used in many applications including helping the organizing the search results, aiding the users in navigating newsgroup, mining the features of participators, investigating hot topics and their evolutions for some period and so on.

We provided a simple example in the experiments, and other applications of PostRank could be explored and developed. Our method provides an essential and simple way to determine the PostRank with link analysis on Usenet. Some improvements can be done to revise it or adjust the parameters according to the requirements.

On the WWW, hyperlinks indicate the choice of web page creators. It has been confirmed that link carries less noisy information than text, and the effectiveness of link analysis has been testified by some web search engines. Similar with web structure, the structure of newsgroup forms gradually along with the evolution of newsgroup. It represents the judgments and choices of participators and reflects the swarm intelligence of the newsgroup. Therefore, it could provide rich helpful information for the task of data mining on Usenet. Together with the IR methods based on text contents, link analysis can be used in the clustering, topic discovery, etc., to make full use of the rich resources on Usenet and other discussion systems.

References

1. Baeza-Yates, R.A., Ribeiro-Neto, B.A.: Modern Information Retrieval. ACM Press/ Addison-Wesley (1999)
2. Brin, S., Page, L., Motwanl, R., Winogard, T.: The pagerank citation ranking: Bring order to the web. Technical report, Stanford University, (1999), Available at `http://dbpubs.stanford.edu:8090/pub/1999-66`
3. Kleinberg, J.: Authoritative sources in a hyperlinked environment. Journal of the ACM 46(5), 604–632 (1999)
4. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proc. of The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, ACM Press, New York (2001)
5. Schwartz, M.F., Wood, D.C.M.: Discovering shared interests using graph analysis. Communications of the ACM 36(8), 78–89 (1993)
6. Tuulos, V.H., Tirri, H.: Combining topic models and social networks for chat data mining. In: Proc. of 2004 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 206–213 (2004)

7. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
8. Xi, W., Lind, J., Brill, E.: Learning effective ranking functions for newsgroup search. In: Proc. of the 27st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, ACM Press, New York (2004)
9. Borgs, C., Chayes, J.T., Mahdian, M., Saberi, A.: Exploring the community structure of newsgroups. In: Proc. of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 783–787 (2004)
10. Agrawal, S.D., Rajagopaian, S., Srikant, R., Xu, Y.: Mining newsgroups using networks arising from social behavior. In: Proc. of the Twelfth International World Wide Web Conference, New York, ACM Press, New York (2003)
11. w3.org: Network news transfer protocol (Internet(WWW)), `http://www.w3.org/Protocols/rfc977/rfc977`
12. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Reviews of Modern Physics 74, 47–97 (2002)
13. cpan.org: Comprehensive perl archive network (Internet(WWW)), `http://www.cpan.org`
14. Langville, A.N., Meyer, C.D.: Deeper inside pagerank. Internet Mathmatics 1, 335–380 (2003)
15. Tsaparas, P.: Link Analysis Ranking. PhD thesis, University of Toronto (2004)
16. Ng, A.Y., Zheng, A.X., Jordan, M.I.: Link analysis, eigenvectors, and stability. In: Proc. of International Joint Conference on Artificial Intelligence (2001)