Learning Information Diffusion Process on the Web

Xiaojun Wan, Jianwu Yang
Institute of Computer Science and Technology, Peking University, Beijing 100871, China
{wanxiaojun, yangjianwu}@icst.pku.edu.cn

ABSTRACT

Many text documents on the Web are not originally created but forwarded or copied from other source documents. The phenomenon of document forwarding or transmission between various web sites is denoted as Web information diffusion. This paper focuses on mining information diffusion processes for specific topics on the Web. A novel system called LIDPW is proposed to address this problem using matching learning techniques. The source site and source document of each document are identified and the diffusion process composed of a sequence of diffusion relationships is visually presented to users. The effectiveness of LIDPW is validated on a real data set. A preliminary user study is performed and the results show that LIDPW does benefit users to monitor the information diffusion process of a specific topic, and aid them to discover the diffusion start and diffusion center of the topic.

Categories and Subject Descriptors:

H.4.m [**Information Systems**]: Miscellaneous

General Terms: Design, Experimentation, Performance

Keywords: Information diffusion, Information flow, Web mining

1. INTRODUCTION

The amount of documents on the Web, including news pages, forum postings, blog articles, etc., has grown exponentially in recent years. However, many documents are not originally created but forwarded or copied from some documents in different web sites. In other words, documents are diffused or transmitted between web sites frequently. Some documents are directly copied or forwarded from one web site to another web site without any changes, and other documents are forwarded between web sites after minor revisions, e.g., addition or deletion of some texts, or rewriting of some sentences. According to our pilot study, more than eighty percent of news documents on popular Chinese news portals (e.g. *sina.com*, *sohu.com*, etc.) are forwarded from other web sites. More than thirty percent of forum postings on popular Chinese forums (e.g. *smth.com*, *bbs.163.com*, etc.) are not originally created.

A specific topic is usually represented by a group of documents sharing the same topic. For example, a hot news topic about "APEC2006" is composed of a set of documents about stories of APEC2006. The topics can be obtained by using topic detection or clustering methods. Likewise, there exist many diffusion relationships between the documents within a topic, and the diffusion process for the topic is represented by all the diffusion relationships. Analyzing and modeling the whole information diffusion behavior is from a macroscopic perspective and it can reveal the underlying mechanism of the Web social network. While identifying the information diffusion process for a specific topic can benefit users to better understand the topic from a microscopic perspective. To the best of our knowledge, most previous works about information diffusion try to model the behavior of information flow on the Web from a macroscopic perspective [3, 4, 5]. More recently, some research works study the dynamics of information propagation in Web blogs, such as [1, 2]. In this paper, we propose LIDPW to identify the information diffusion process for a given topic. The task of information diffusion process identification can be formally defined as follows:

Given a set of documents $D=\{d_1,d_2,\ldots,d_n\}$ belonging to a specific topic, each document d_i is associated with a tuple (t_i, $LocationSite_i$), where t_i is the timestamp denoting the time when the document is published and $LocationSite_i$ is the name of its current web site. LIDPW aims to find another tuple ($SourceSite_i$, $SourceDoc_i$) for d_i, where $SourceDoc_i$ is the source document that document d_i is copied or forwarded from, and $SourceSite_i$ is the location site of $SourceDoc_i$, in other words, $SourceSite_i$ is the site from where document d_i is copied or forwarded from. A typical document diffusion process can be represented by [$LocationSite_j$: d_j—>$LocationSite_i$: d_i] ($t_j<t_i$), which means that document d_i located in $LocationSite_i$ is forwarded or copied from document d_j located in $LocationSite_j$. Thus we have $SourceSite_i = LocationSite_j$ and $SourceDoc_i=d_j$.

The proposed LIDPW identifies information diffusion process by the following steps[1]:

Document Metadata Extraction: For a specified document d_i, its metadata mainly includes the timestamp t_i and the name of its location site $LocationSite_i$. The timestamp is extracted by a simple time extraction tool which can identify the timestamp in a web page. If more than one timestamp is extracted, the timestamp in the middle of the web page will be selected as the final timestamp of the document. For those regularly designed web pages from popular news portals, they share almost the same template for arranging the elements, including source site, we can use simple templates to extract the source sites of the web pages as another type of metadata if available.

Document Sorting: The documents are sorted by the timestamp. The documents published earlier are placed at the first of the list.

Diffusion Process Identification: This step aims to identify document diffusion relationships one-by-one from the ordered list using machine learning techniques. The process of identifying diffusion process is iterated on each document in the ordered list from the first document to the last document. For document d_i and each preceding document d_j ($t_j<t_i$) in the list, the problem of whether d_j is the source document of d_i can be solved by a binary classification algorithm. In this study, the Support Vector Machine implemented in the SVM-Light toolkit is adopted and the following features are computed and fed into the learning algorithm:

[1] Though we focus on Chinese Web documents in this study, the proposed algorithm is deemed to be language-independent.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

Metadata-Based Features: We use one feature to indicate whether the extracted source site of d_i is available, and if available whether the extracted source site of d_i is equal to the location site of d_j;

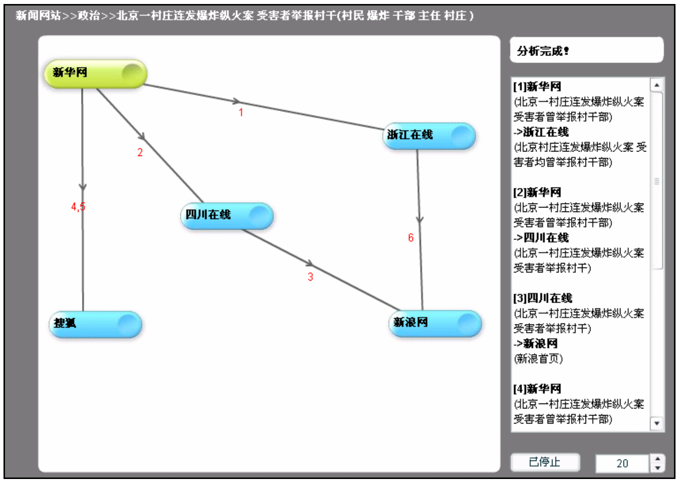
Cueword-Based Features: We define a list of 16 Chinese cue words, namely the variant forms of "forward", "from" or "source". For each word, the corresponding feature is to indicate whether the name of the location site of d_j appears and follows the cue word in the prefix or suffix string of the text of d_i. The length of the prefix or suffix strings is heuristically set to 50 characters.

Similarity-Based Features: We compute the Cosine similarity value between d_i and d_j, and then use the similarity value as one feature; in addition, we define a feature to indicate whether the value is the largest similarity value between d_i and all preceding documents in the list.

If more than one preceding document is classified as the source document of d_i, we use the document with the highest confidence value returned by the SVM-Light as its final source document. The source site of d_i is the location site of the source document.

Result Demonstration: After all the diffusion relationships are identified for the given topic, LIDPW further deduces the start site and the center site of the diffusion process. The start site is the web site where the document is created most originally. The center site is the web site which forwards or copies maximum documents to other sites. The start site and the center site are of great important for users to monitor the diffusion process.

Lastly, LIDPW dynamically presents all the relationships on a demo one-by-one. On the demo, the start site and the center site are marked in different color. Figure 1 shows an example of diffusion process and the start site and center site are the same site at the top-left corner.



新闻网站>>政治>>北京一村庄连发爆炸纵火案 受害者举报村干(村民 爆炸 干部 主任 村庄)

分析完成!

[1]新华网
(北京一村庄连发爆炸纵火案受害者曾举报村干部)
->浙江在线
(北京村庄连发爆炸纵火案 受害者均曾举报村干部)

[2]新华网
(北京一村庄连发爆炸纵火案受害者曾举报村干部)
->四川在线
(北京一村庄连发爆炸纵火案受害者举报村干)

[3]四川在线
(北京一村庄连发爆炸纵火案受害者举报村干)
->新浪网
(新浪首页)

[4]新华网
(北京一村庄连发爆炸纵火案受害者曾举报村干部)

Figure 1. An example of diffusion process in LIDPW

2. EVALUATION

We manually labeled the diffusion processes for 30 recent topics (15 news topics and 15 forum topics) produced by our in-house system for Chinese Web topic detection, in which 20 topics were used for training and the other 10 topics were used for test. Table 1 gives the classification results for different feature sets. The results are averaged over all diffusion relationships across topics.

Seen from Table 1, the metadata-based feature set achieves the best precision value, which shows that the metadata is the most convincing evidence for accurately identifying the source document or site. Both the cueword-based feature set and the similarity-based feature set can benefit to find source documents or sites for the documents with no explicit tags, thus improving the recall value. The best overall performance (F-measure) is achieved based on all the features.

Table 1. Performance comparison

	Metadata	Metadata+Cueword	Metadata+Cueword+Similarity
Precision	**0.93**	0.77	0.75
Recall	0.51	0.80	**0.86**
F-measure	0.66	0.78	**0.80**

In order to evaluate the whole system from users' perspective, we performed a preliminary user study. The user study involved with 10 subjects (users) and each subject was requested to fill in a questionnaire after using the system. The questionnaire contained five questions involving with effectiveness, real-time, user interface, usefulness and overall satisfaction. Subjects were required to express an opinion over a 5-point scale for each of the questions, where 1 stood for "*not at all*", 3 for "*somewhat*" and 5 for "*extremely*". We collected the responses of subjects and averaged them, as shown in Table 2. Seen from the table, the users thought the effectiveness of the system is ok and the system is pretty real-time. They liked the user interface very much and most of them considered the system was useful in practice. Overall, most of them were satisfied with the system.

Table 2. Results of user study

	Effectiveness	Real-time	Interface	Usefulness	Satisfaction
User 1	3	5	4	3	4
User 2	3	4	2	4	3
User 3	4	5	5	4	4
User 4	4	5	3	3	4
User 5	4	4	5	3	4
User 6	3	5	4	4	3
User 7	2	3	5	3	3
User 8	2	5	4	3	2
User 9	4	4	4	5	5
User 10	4	3	5	4	4
Average	3.3	4.3	4.1	3.6	3.6

3. ACKNOWLEDGMENTS

We thank Dong Wang, Jun Yu and Xiaojiang Huang for implementing the demo system. This work was supported by the National Science Foundation of China (60642001).

4. REFERENCES

[1] D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins. Information diffusion through blogspace. In Proceedings of WWW2004.

[2] R. Kumar, J. Novak, P. Raghavan, A. Tomkins. On the bursty evolution of blogspace. In Proceedings of WWW2003.

[3] C. Moore, M. E. J. Newman. Epidemics and percolation in small-world networks. Phys. Rev. E, 2000, 61:5678-5682.

[4] R. Pasto-Satorras, A. Vespignani. Epidemic spreading in scale-free networks. Phys. Rev. Letters, 2001, 86(14):3200-3203.

[5] D. Watts, S. Strogatz. Collective dynamics of 'small-world' networks. Nature, 1998, 393:440-442.