Journal of Southeast University (English Edition) Vol. 24, No. 3, pp. 381 – 384 Sept. 2008 ISSN 1003—7985

System of twice-gathering information and research of information fingerprint HashTrie

Shen Yang[1] Zhu Chanyuan[2] Li Shuchen[3]

([1] School of Information Management, Wuhan University, Wuhan 430072, China)
([2] School of Computer Science, Wuhan University, Wuhan 430072, China)
([3] International School of Software, Wuhan University, Wuhan 430072, China)

Abstract: This paper presents a twice-gathering information interactive system prototype of e-government based on the condition that the Intranet and the Extranet are physical isolated. Users in the Extranet can gather links of the latest related information from client software which is previously collected by web alert in the Internet. Finally, through ferry-type transport devices, information is browsed by users in the Intranet, and it is transported to a storage device and synchronized with the web platform in the Intranet. During information gathering in the Extranet and data synchronization in the Intranet, it is essential to avoid repeated gathering and copying by means of comparing the extracted information fingerprints gathered from the web pages. This prototype uses HashTrie to store information fingerprints. During testing, the structure based on HashTrie is 2.28 times faster than the Darts (double array Trie) which is the fastest structure in the existing applied patent. The existing 12 types of high speed Hash functions serving for HashTrie are also implemented. When the dictionary content is larger than 5×10^5 words, the PJWHash or the SuperFastHush function can be adopted; when the dictionary content is 10^5 words, CalcStrCR32 and ELFHash functions can be adopted.

Key words: physical isolation; twice-gathering; duplicated web pages elimination; information fingerprint; HashTrie

Physical isolation is considered as a means of security management, which effectively keeps away any network threat from the outside to the Intranet. It forbids all the data exchanges between the network information system and any suspected outside websites to further guarantee the security of the Intranet and the information system. On the one hand, physical isolation ensures the security of the Intranet[1-2]. On the other hand, because physical isolation separates the network and forbids data exchanges, users cannot communicate and interact with the outside in real-time, and thus become an "information-isolated island". As several operation forms in e-government are highly demanding both in network security and information interaction, users' requests are increasing so that data can be exchanged within limitations between physical isolation networks regarding the requirements of operation and application[3] and users can acquire information from the Extranet to the Intranet swiftly. Governments, the military, financial and security institutions require a product of information gathering and interaction which can sufficiently guarantee self-security as well as implement network communication[4]. Now, several information interaction systems based on physical isolation have come on the market, such as CopGap, but they have some disadvantages in that they cannot collect information which users are interested in on an unknown website, and they repeatedly gather similar web pages. On the subject of dynamic gathering, Google web alert can automatically trace and report information which matches user interests, but there is little information gathering use of this technology based on physical isolation. On the subject of avoiding collecting repeated web pages[5-6], Fogaras et al. [7] used information fingerprints which are stored in the content structure tree to assess similarities of web pages and eliminate repetition ones. However, when the number of web pages is great, efficiency decreases sharply. Monika[8] proposed a new algorithm that is combined with another algorithm which prevails among search engines in finding similar web pages to solve the problem when faced with a large number of web pages. Google Blog extracts information fingerprints from URLs to eliminate redundant web pages, yet there is much space for web page gathering technology, proposed by Larbin and Nutch, to improve the speed of extracting information fingerprints.

1 Twice-Gathering Information: Designing Idea and Implement Approach

To gather integrate information on condition that users are separated from the Extranet, users obtain data they need by means of gathering and twice-gathering. Gathering means that users find information by information source addresses provided by users. While twice-gathering means that it is based on web alert technology to collect, subscribe, and trace information, and then find information by information source addresses. Twice-gathering is based on whole-network coverage of the search engine. It effectively collects information from unknown websites which may interest users.

Web alert is an automatic web searching service which is provided by the search engine. It helps corporations and users see information they are interested in from the latest web pages in the Extranet. The searching results of the latest related web pages are sent to users who subscribe to this service by email, daily or weekly. Twice-gathering is a process which uses web alert's feedback of information source addresses from the email to find the actual information and sends it to users. First-gathering and twice-gathering ensure that the feedbacks sent to users are integrated information re-

Received 2008-04-15.
Biography: Shen Yang (1974—), male, doctor, associate professor, 124739259@qq.com.
Foundation item: The National Basic Research Program of China (973 Program) (No. 2007CB310806).
Citation: Shen Yang, Zhu Chanyuan, Li Shuchen. Systems of twice-gathering information and research of information fingerprint HashTrie[J]. Journal of Southeast University (English Edition), 2008, 24(3): 381 – 384.

sources rather than linking addresses.

The design approach of this prototype is as follows: In the environment of physical isolation between the Intranet and the Extranet by isolation devices, users in the Intranet employ a certain information configuration interface to set information searching conditions and requirements. The automatic agent and the searching system analyze and integrate user configuration requirements, which collect, subscribe, and trace information from the Extranet and finish information gathering and twice-gathering. Then, the search results sent by transport devices come to a storage device in the Intranet and synchronizes the data on the Intranet network platform so that users in the Intranet can browse it. Meanwhile, the agent system can transport information which is authorized by the Intranet user to the Extranet with information security and synchronization between the Intranet and the Extranet.

To sum up, the particular process of information interaction of this prototype is as follows. First, construct an Intranet network platform that users in the Intranet can visit and browse through. Secondly, according to user operation and implement requirements, configure information searching terms, with personalized conditions and requirements, which includes parameters of information sources, information types, information validity, information update times, etc. Thirdly, use automatic agents and searching systems to collect, subscribe to, and trace information from the Extranet, and gather information through gathering and twice-gathering. Fourthly, do security-checks with interactive information. Finally, searching results are sent to network storage devices through transportation devices. After restarting, switch off the Extranet, and synchronize the data between the storage devices and the Intranet network platform, so that users can search and browse them directly. Meanwhile, transportation devices indicate all the types of portable storage media including mobile disks and u-disks. Its function is to load data packages, including integrated data, increment data, and differential data, which are going to be exchanged between the Intranet and the Extranet. Then, data are synchronized between the Intranet and the Extranet by means of artificial transportation. At the same time, the agent system can upload information which is authorized by users from the Intranet to the Extranet, so that information interaction succeeds with a high security guarantee provided by physical isolation. In the meantime, the Intranet user authority information indicates those who are permitted to upload and release by users. Users use privacy configuration interfaces to set information grades. The agent system leaches information which lacks releasing authority according to user configurations automatically.

2 Information Fingerprint Extraction

When collecting information from the Extranet and synchronizing information in the Intranet, extracting and comparing information fingerprints from web pages to prevent repeated gathering and copy is needed[9–10]. That is the pivotal technology in this prototype. The prototype adopts HashTrie to save information fingerprints and updates times of web pages. Fingerprints mean characterizations which are extracted from information, commonly including an array of words or an array of words as well as their weights. Then the array of words calls a Hash function to transform it into an array of numeric values. This array of values marks a piece of information as its fingerprint. In theory, any two characterizations of different texts are different, so the array of numeric values should differ from each other, like the fingerprints of humans. In the system of twice-gathering, it is required to extract information fingerprints from both URLs and texts in web pages.

To give an example of extracting fingerprints from texts in web pages by this prototype: First, use the ROST TextExtractor module, which is developed by the authors, to extract text in web pages. After that, use the ROST WordParser module to parse words in the extracted text. Then, apply the ROST TFIDF module to analyze and extract the characterization words of a web page. Moreover, construct a vector array with these characterization words, and the top N characterization words are regarded as an information vector, e. g. [search 15, engine 10, SEO 6]. Next, call a Hash function directly for information fingerprints and make use of Boolean comparison. All three modules mentioned above are independent software applications, and you can obtain them for free by googling them.

Most of the literature uses the Hash table to extract information fingerprints. According to the authors' related studies on Chinese wordparsing, the speed of the Hash table is a third or less than that of HashTrie which applies the same Hash functions, so it directly uses HashTrie to store and analyze information fingerprints from web pages.

HashTrie is a new effective data structure which combines the capabilities of the Hash table and the Trie tree. Compared to the original Hash table, the nodes of HashTrie are not fixed, and each node is an array of numbers. In general, the length of each array is 256, and there is little difference between the initial HashTrie and the original Hash table. Resolution of the conflict is to employ an overflow table of the self-similar HashTrie structures. We can call all types of Hash functions in HashTrie, and the output results of the Hash function has to be a 32 bit unsigned integer as an address pointer. We use a low byte to address in the table and set a quaternion to save other data. HashTrie is defined as follows:

```
type {THashTrie}
 PPHashItem = ^PHashItem;    //The start pointer of HashTrie.
 PHashItem = ^THashItem;    //The pointer of each Hash item.
THashItem = record
Next: PHashItem   //The Hash pointer of next item.
Key: widestring   //The key.
Value: pointer     //Store data.
Freq: integer     //Frequency.
end
```

3 Related Experiments

Thirteen different Hash functions are implemented, and ROSTHash is raised. When polling 0.1 million keywords, like "刻舟求剑", respectively, 1 million times, in dictionaries of 0.56 million words and in dictionaries of 0.1 million words, the following results can be obtained in Tab. 1. Test computer: P4-3GHz CPU, 512MB RAM, Windows XP OS; unit: s; order by: three fields.

It can be seen from the above results that the speeds of various Hash functions differ. Speed in polling a certain word may not be swift in polling in a 0.1 million words dictionary, while speed in polling in a 0.1 million words dictionary may not be fast in a 0.56 million words dictionary. Considering the comprehensive capacity, we can neatly apply HashTrie with PJWHash, SuperFastHash, ELFHash, and CalcStrCRC32 in it. In this experiment, HashTrie implements a high-speed count of words, so that a high speed rechecking capacity of web pages is implemented. We compare it with an invention patent (Perfect Double Array TRIE Tree Dictionary Management No. 200510130690.3). According to the description provided by the applicant, PDATT can poll 1 million words per second under the same computer conditions. The speed of the proposed algorithm which applies the fastest PJWHash function etc. as well as information fingerprint searching is 2.28 times faster than that of the patent.

Tab. 1 The speed test of Hash functions

Hash function	0.56 million	0.1 million	1 million times
PJWHash	0.093	0.063	0.437
SuperFastHash	0.093	0.078	0.782
ELFHash	0.108	0.046	0.421
DJBHash	0.108	0.109	0.437
RSHash	0.109	0.108	0.421
BKDRHash	0.296	0.079	0.437
CalcStrCRC32	0.437	0.046	0.765
APHash	0.469	0.078	1.375
JSHash	0.61	0.093	1.421
ROSThash	0.625	0.078	0.781
FNVHash	0.656	0.125	0.781
DEKHash	0.764	0.108	0.781
SDBMHash	0.782	0.140	0.796

4 Conclusion

We propose a prototype of an e-government information gathering interactive system based on the conditions of physical isolation. That means, users in the Intranet can communicate with the Extranet for information synchronization and interaction through isolation systems, automatic agents and searching systems, transport devices, storage devices, etc. It can be applied to the departments and fields which simultaneously need information security and information interaction, such as governments, financial institutions, the military, e-governments, e-businesses, etc. In this research, twice-gathering information is a means of information gathering which has never appeared in other literature up to the present. To avoid repeated gathering and copying, this prototype uses HashTrie to store information fingerprints, and to pursue the best speed, we present a new type of Hash function which applies the existing 12 types of high speed Hash functions. After assessment and comparison, it is concluded that when a dictionary content is larger than 5×10^5 words, PJWHash or SuperFastHush functions can be adopted; when a dictionary content is 10^5 words, CalcStrCR32 and ELFHash functions can be adopted. Since the structure based on HashTrie is 2.28 times faster than Darts, which is the highest speed in applied patents at present, we will continue researching on information fingerprint of spanned languages, and the synchronization of twice-gathering information in portable devices.

References

[1] Garmeli B. *Gap appliance enhance security* [M]. Network World, 2001: 36 – 39.

[2] Northcutt S, Zeltzer L, Winters S, et al. *Inside network perimeter security* [M]. New Riders, 2003: 78 – 83.

[3] Zhang Sida. Information ferrying system [J]. *Journal of Chengdu University of Information Technology*, 2004, **19**(1): 62 – 65. (in Chinese)

[4] Soumen Chakrabarti. *Mining the web* [M]. San Francisco: Morgan Kaufmann Publishers, 2003: 118 – 120.

[5] Yan T W, Garcia-Molina H. Duplicate removal in information dissemination [C]//*Proceedings of the* 21*st International Conference on Very Large Data Bases* (*VLDB*'95). San Francisco, CA, USA, 1995: 66 – 77.

[6] Di Iorio E, Diligenti M, Gori M, et al. Detecting near-replicas on the web by content and hyperlink analysis [C]//*Proceedings of the IEEE/WIC International Conference on Web Intelligence*. New York: IEEE Computer Society Press, 2003: 249 – 255.

[7] Fogaras D, Racz B. Practical algorithms and lower bounds for similarity search in massive graphs [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2007, **19**(5): 585 – 598.

[8] Monika H. Finding near-duplicate web pages: a large-scale evaluation of algorithms [C]//*Proceedings of the* 29*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Seattle, 2006: 284 – 291.

[9] Dean J, Henzinger M R. Finding related pages in the world wide web [J]. *Computer Networks*, 1999, **31**(11): 1467 – 1479.

[10] Haveliwala T, Gionis A, Klein D, et al. Evaluating strategies for similarity search on the web [C]//*Proceedings of the* 11*th International World Wide Web Conference*. Hawaii, USA, 2002: 432 – 442.

二次信息采集系统及信息指纹 HashTrie 研究

沈　阳[1]　朱婵元[2]　李舒晨[3]

([1] 武汉大学信息管理学院,武汉 430072)
([2] 武汉大学计算机学院,武汉 430072)
([3] 武汉大学国际软件学院,武汉 430072)

摘要:提出一种在内网和外网间处于物理隔离状态下防止信息重复采集的电子政务二次信息采集交互系统原型. 外网用户能够从客户端软件中二次采集由 web alert 功能采集的互联网中最新相关网页的链接所指内容,最后再通过摆渡式传输设备将采集结果传递到存储设备上,与内网搭建的网络平台进行数据同步,供内网用户直接浏览. 在外网抓取信息和内外网数据同步中,都需要对网页提取信息指纹进行对比,防止重复抓取和拷贝. 原型采用 HashTrie 保存信息指纹. 进行评测对比后,可知基于 HashTrie 信息指纹提取比目前专利申请中速度最快的 Darts(双数组 Trie)结构快 2. 28 倍,还提出了一种新的 Hash 函数,并且实现了现有 12 种高速 Hash 函数以供 HashTrie 使用,当词典容量大于 50 万词时,可以采用 PJWHash 或 SuperFastHash 函数,而当词典容量为 10 万词时,可以采用 CalcStrCRC32 和 ELFHash 函数.

关键词:物理隔离;二次抓取;网页去重;信息指纹;HashTrie
中图分类号:TP393. 09