Journal of Southeast University (English Edition) Vol. 22, No. 3, pp. 423 – 425 Sept. 2006 ISSN 1003—7985

Algorithms of mining data records from website automatically

Qiu Yong Lan Yongjie

(School of Information and Electronic Engineering, Shandong Institute of Business and Technology, Yantai 264005, China)

Abstract: In order to improve the accuracy and integrality of mining data records from the web, the concepts of isomorphic page and directory page and three algorithms are proposed. An isomorphic web page is a set of web pages that have uniform structure, only differing in main information. A web page which contains many links that link to isomorphic web pages is called a directory page. Algorithm 1 can find directory web pages in a web using adjacent links similar analysis method. It first sorts the link, and then counts the links in each directory. If the count is greater than a given valve then finds the similar sub-page links in the directory and gives the results. A function for an isomorphic web page judgment is also proposed. Algorithm 2 can mine data records from an isomorphic page using a noise information filter. It is based on the fact that the noise information is the same in two isomorphic pages, only the main information is different. Algorithm 3 can mine data records from an entire website using the technology of spider. The experiment shows that the proposed algorithms can mine data records more intactly than the existing algorithms. Mining data records from isomorphic pages is an efficient method.

Key words: data mining; data record; website; isomorphic page

There are many important data on the web. These data are as mineral resources distributed among many websites. How to mine certain "mineral resources" from many websites and bring them together to abstract the useful data is very significant[1–2]. But considering the impressive variety of the web, retrieving interesting content has become a very difficult task. Web content mining uses the ideas and principles of data mining and knowledge discovery to screen more specific data[3–4]. Another important aspect of web content mining is the usage of the web as a data source for knowledge discovery. This offers new interesting opportunities since more and more information regarding various topics is available in it[5–6]. But the use of the web as a provider of information is unfortunately more complex than working with static databases. Because of its very dynamic nature and its vast number of documents, there is a need for new solutions that are not dependent on accessing the detailed data at the outset[7–8].

Recently, researchers have been exploring new approaches to fully automate wrapper construction. That is, without user training examples. For example, OMONI[5] and MDR[6]. Both OMONI and MDR can extract information from multi-recording web pages. The main point is to identify repetitive record structure automatically. OMONI[5] is a fully automated object extraction system. It parses web pages into tree structures and performs object extraction. The OMONI object extraction process consists of three phases: ① Preparing a web document for extraction; ② Locating objects of interest in a web page; ③ Extracting objects of interest in a page. Phase ③ consists of two tasks: candidate object construction and object extraction refinement. Candidate object construction is the process of extracting objects from the raw text data of the web document using the object separator tag identified in phase ②. After the object separator tag is chosen, the objects need to be extracted from the components of the chosen subtree. OMONI has a serious disadvantage: It is able to mine contiguous data records only, and cannot mine noncontiguous data records. The "non-contiguous data records" means that two or more data records intertwine in terms of their HTML codes. To solve the problem of mining "non-contiguous data records", MDR was proposed in Ref. [7]. In MDR, the regularly structured information in the web is called data records. MDR can mine data records in a web page automatically. It currently finds all data records formed by table and form related tags, i. e. , table, form, tr, td, etc. MDR can mine both contiguous and noncontiguous data records.

However, web pages that maintain multi-records are actually directory pages. The information in a directory pages is not detailed, the detailed information exists in a lower-level web page which has single re-

Received 2006-04-25.

Biography: Qiu Yong (1959—), male, professor, sdytqy@163.com.

cord information only, so it cannot be extracted using a duplicated-record finding algorithm. To solve this problem and extract detailed information from the web, this paper introduces a concept of isomorphic web page, and proposes algorithms to find directory pages and extract main data from isomorphic web pages.

1 Mining Data Record from Website Automatically

As described above, OMONI[6] and MDR[7] can extract information from multi-recording web pages automatically. The main point is to identify repetitive record structures automatically. However, web pages maintaining multi-records are actually directory pages. The information in directory pages is not detailed. The detailed information exists in lower-level web pages, called detailed pages. This situation always occurs in e-Commerce or news web sites. A detailed page has only one record of information, so it cannot be extracted using duplicated-record finding algorithms. To solve this problem, two algorithms are proposed, one algorithm for finding a directory page, the other for extracting the main information from isomorphic web pages.

Definition 1 Directory web page: A web page containing many links that link to isomorphic web pages.

Definition 2 Noise information: The advertisements, links and contacts etc. are in web pages.

Definition 3 Main information: In an isomorphic web page, information is separated from noise information.

Definition 4 Isomorphic web page: A set of web pages that have uniform structure, differing only in main information. There are many isomorphic web pages in e-Commerce or new websites.

To extract detailed information from web pages, we propose a novel algorithm IDF to find isomorphic web page directories in directory pages, propose an algorithm IJ to judge the degree of two isomorphic web pages, and an algorithm MI for main information mining from isomorphic web pages.

Function 1 IJ: function for isomorphic web page judgement.

Input: web page P1, P2.
Return: Percent of isomorphic degree.
Function IWDF (web page P1, web page P2)
{For $i = 1$ to min(p1. tagscount/2, p2. tagscount/2), do
If P1. tags[i]. tagname < > P2. tags[i]. tagname then break; $T = i$;
For $i = 1$ to min(p1. tagscount/2, p2. tagscount/2) − 1, do
If P1. tags[p1. tagscount − $i + 1$]. tagname < > P2. tags[$i + 1$]. tagname then break;
Return ($T + i$)/max(p1. tagscount, p2. tagscount) 100 }

Algorithm 1 IDF isomorphic web page directory finding algorithm.

Input: Web page W, count threshold (CT), isomorphic threshold (IT).
Output: Isomorphic web page directory D.
Procedure IDF (W, CT, IT)
{Isomorphic web page directory finding algorithm
Sort link directory + serial number
For adjacent link directory, accounting each directory
If the count of identical directory(d) > CT, then $D = D \cup \{d\}$
For all $d \in D$
get S1, S2 from web page where S1 and S2 have the directory of d.
If IJ(S1, S2) < IT, then $D = D - \{d\}$
If D has multi directory, then
Sort D according to its tag number of web page, take the directory the web page tag number maximal, delete the others. }

Algorithm 2 MI: algorithm of main information mining from isomorphic web page.

Input: Isomorphic web page P1, P2.
Output: Offset of main information first tag from top: TOffset,
Offset of main information last tag from bottom: BOffset,
Procedure MI (web page P1, web page P2)
{ $N = 0$;
While(P1. tags[n]. tagname = P2. tags[n]. tagname And P1. tags [n]. information = P2. tags[n]. information)
$N = N + 1$; Toffset = $N - 1$; $N = 0$;
While(P1. tags[P1. tagscount-n]. tagname = P2. tags[P2. tagscount-n]. tagname And P1. tags[P1. tagscount-n]. information = P2. tags[P2. tagscount-n]. information)
$N = N + 1$; Boffset = $N - 1$; }

Algorithm 3 MDRA: algorithm of mining data record from website automatically.

Procedure MDRA(website W)
{For all page w in W do
{ D = NULL; IDF (w, CT, IT)/ * Find directory page * /
If D is not Null then
MI($d1$, $d2$); / * $d1 \in D$, $d2 \in D$, Find TOffset and BOffset * /
For all d in D Do extract tags data from TOffset to BOffset in d;
}

Algorithm MI is based on three observations: ① A detailed isomorphic webpage has only one type of information, called main information; ② Main information in isomorphic web page is in the middle of the webpage; ③ The noise information before the main information and after the main information in two isomorphic webpages is the same, only the main information is different.

2 Experiment

In our experiment, we evaluate the effectiveness of our method. We download and cache web pages from 20 different web sites. Our system is implemented in Visual C + +. All the experiments are conducted on

a Pentium 4, 1.8 GHz PC with 512 MB RAM. The results are shown in Tab. 1.

Tab. 1 Experimental results

URL	Isomor-phic pages number	CT	IT	Identified pages	Time/ms
chemstore. cambridge-soft. com	5	3	65	5	656
www. godaddy. com	4	3	65	6	453
www. compusa. com	8	3	65	8	876
www. radioshack. com	9	3	65	9	1 210
www. earlemu. com	4	3	65	4	563
www. kadybooks. com	20	3	65	25	2 410
www. kidsfootlocker. com	9	3	65	9	1 167
shop. lycos. com	13	3	65	13	1 425
thenew. hp. com	4	3	65	6	454
www. dell. com	5	3	65	5	655

The experimental results show that the proposed algorithms can give perfect results for every page.

3 Conclusion

In this paper we have discussed the problem of mining data records from websites automatically and suggested a detailed information mining approach for solving it. Taking into consideration the limitations of the existing method, our method can extract more detailed information from web pages. The proposed algorithms have been used in the development of "a large e-commerce navigation site support platform"(ECSP). The ECSP is a web mining based system. It has functions such as: discovery of web topic contents, downloads of web topic contents, automatically analysis and deeply collect in web, data center, collect examine edit, information classification, misty recognition, automatic fault-tolerance, information filtering, the dynamic web page, information announcements.

References

[1] Amitay E, Paris C. Automatically summarizing web sites: is there a way around it?[A]. In: *Proc of the* 9*th International Conference on Information and Knowledge Management* [C]. New York: ACM Press, 2000. 173 - 179.

[2] Cohen W, McCallum A, Quass D. Learning to understand the web[J]. *IEEE Data Engineering Bulletin*, 2000, **23**(3): 17 - 24.

[3] Embley D, Jiang Y, Ng Y. Record-boundary discovery in web documents[A]. In: *Proc of SIGMOD*[C]. Philadelphia, 1999. 213 - 219.

[4] Han J, Chang K C C. Data mining for web intelligence [J]. *IEEE Computer*, 2003, **10**(5): 51 - 62.

[5] Buttler D, Liu L, Pu C. A fully automated extraction system for the world wide web[A]. In: *Proc of the* 21*st International Conference on Distributed Computing* [C]. Phoenix: IEEE Press, 2003. 361 - 370.

[6] Liu B, Grossman R, Zhai Y. Mining data records in web pages[J]. *UIC Technical Report*, 2004, **5**(1): 35 - 47.

[7] Zaki Mohammed J. Efficiently mining frequent trees in a forest: algorithms and applications [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2005, **17**(8): 516 - 527.

[8] Sun J T, Zeng H J, Liu H, et al. Cubesvd: a novel approach to personalized web search[A]. In: *Proceedings of the* 14*th International Conference on World Wide Web* [C]. New York: ACM Press, 2005. 652 - 662.

从网站中自动挖掘数据记录的算法

邱 勇 兰永杰

(山东工商学院信息与电子工程学院,烟台 264005)

摘要:为了提高从 web 中挖掘数据记录的精确性和完整性,提出了同构页与目录页的概念及 3 个算法. 如果一组网页结构相同,只是主信息不同,该网页称为同构页. 一个包含有多个指向同构页连接的网页称为目录页. 算法 1 用于发现目录页,它首先将连接排序,并对同一目录的链接记数,如果记数大于某一给定阀值,则对其链接子页进行相似比较并得到结果. 同时给出了一个网页相似度判断的函数. 算法 2 采用了噪声信息过滤方法从同构页中挖掘主信息并得到数据记录,该算法是基于在 2 个同构页中噪声信息相同而只有主信息不同. 算法 3 通过采用 Spider 技术可以实现从整个网站中自动挖掘数据记录. 实验表明所提算法比已有算法可挖掘更完整的数据记录. 从同构页中挖掘数据记录是一种有效的方法.

关键词:数据挖掘;数据记录;网站;同构网页

中图分类号:TP311