A DOM-based Anchor-Hop-T Method for Web Application Information Extraction

Yuanyuan Zhang
College of Information Technology, Zhejiang Chinese Medical University, Hangzhou 310053, China
Email: zyy@zjtcm.net

Qinyan Zhang *
Computer Center, Zhejiang University, Hangzhou 310058, China
Corresponding Email: zqy_hellen@sina.com.cn

Guanfu Jiang
College of Computer Science, Zhejiang University, Hangzhou 315100, China

***Abstract*—In order to implement the information fusion of electronic products, the widely adopted approach is to extract information from HTML structure of business Website with deeply data processing. However, modeling Web application is hard to be solved that the data in HTML is semi-formal which displayed as DOM (Document Object Model) tree when using XML schema to data analysis. How to understand and extract information is first to be researched. The general model Anchor-Hop considering the text property and label property is simple to handle this problem. Therefore, it has low effectiveness. This model is sensitive to the data of HTML structure, that if the website structure is slightly changed the issue of extraction accuracy is encountered. As a result, the extraction rules should be redefined because of the changed structure. In order to improve extraction efficiency, this paper proposed a DOM-based dynamic model Anchor-Hop-T information extraction model. The HTML tags including table, ol and ul can be searched and processed using XPath so that it is convenience to extract corresponding Anchor data block. Furthermore, the location of Hop point is considered as invariant, by which our new model based on Anchor and Hop point introduces more concepts for extracting information, such as Anchor data block, Anchor locating library and AH relevance value. Finally, we try to give out an experiment to demonstrate the applicability of our approach.**

***Index Terms*—Web Application, Information Extraction, DOM, Anchor-Hop-T Model**

I. Introduction

With increasing presence and adoption of Webpage on Internet, information extraction becomes a hot topic in Web research [1]. From the processed data of discovered information, business enterprise can change its marketing strategy in order to get maximize profits. At the respect of customer, they can use these data to connect or share their experiences, or even discover new friends. To this end, Web information extraction is to extract information using Website as information source, where information from semi-structured expression in Webpage can be organized as the structured display. Many method has been proposed including data mining [2], rule-based method, machine learning method [3], and Ontology-Based [4] method of Web Information Extraction. Mining with association rules can find the association relationship which is hidden among data, and it is the most common thing in the association knowledge discovery [5]. Machine learning methods based on statistical models are mainly Hidden Markov Model (HMM), maximum entropy model (MEM) and maximum entropy Markov model (MEMM) [6, 7]. Rule-based information extraction is a process by which structured objects are extracted from text based on user-defined rules [8]. However, they all have some shortcomings, such that, rule-based method needs user predefined rules so that the exception will not be automatically handled, the construction of Ontology itself is a difficult job.

Facing a massive amount of information on the Web, how to positioning information domain is most important in Web information extraction. Different from above research, this paper focuses on electronic products-oriented information fusion, which needs information extraction of electronic products. As we investigated some big E-business Website in China, electronic products information are always in the form of table, ul, or lo tags, by which the price, name, and picture of electronic products can be displayed clearly. Simply, our goal is to analyze these tag formats and extract electronic product information form commercial Website.

Based on the previous research, Anchor-Hop-I model, this paper surveys the web content label and structure characteristics, and proposes an Anchor-Hop-T model. There are three concepts, Anchor point, Hop point and absolute path matching algorithm. 1) The Anchor point is datum point of relative path, where all extraction should start with offset location of observation point that is compared this point with the Anchor point. 2) Hop is the path of information point relative to Anchor. According to the offset from Anchor, Hop path is the target information which needs to be extracted. This can be implemented by XPath [9]. For example, the statement "/children//ul/li" means that search li tags information of

© 2014 ACADEMY PUBLISHER
doi:10.4304/jsw.9.3.641-647

any ul sub-node under child node. 3) The absolute path matching algorithm gives the accuracy Anchor point of Anchor-Hop model based on the matched offset of the Anchor path in the DOM tree. However, in Anchor-Hop-I model, the changed structure is main problem that it is hard to be solved. While Anchor-Hop-T model we proposed is a better model than Anchor-Hop-I model. In the case of web structure changing, our new model not only avoids the defect of Anchor-Hop-T model but also gains higher efficiency of extraction than using Anchor-Hop-I. Furthermore, it also can automatically generate template library for locating anchor tags, which greatly reduces the costs of artificial maintenance of extraction rules after extraction.

The remainder of this paper is organized as follows. Section 2 gives the concept of Anchor-Hop-T model. Section 3 presents the implementation of Anchor-Hop-T model. Section 4 discusses experiment and result analysis. Section 5 draws a conclusion and discusses future works.

II. The Concept of Anchor-Hop-T Model

The Anchor-Hop-T model is an extension of Anchor-Hop-I model, where the main difference attributes to the way of locating anchor and extraction rules generation. The Anchor-Hop-T model improves the Anchor-Hop-I model without any change of definition Anchor and Hop. However, it introduces the new concepts, including Anchor Data Block, Anchor Locating Library and Correlation Degree of Anchor.

Generally, the body tag is the main framework of the HTML Webpage. And div tags considered as closure container can be conveniently embedded into body domain, that *ul* and *li* tags integrated together is to display products disorderly. As Fig.1 shown, they are interpreted as follows:

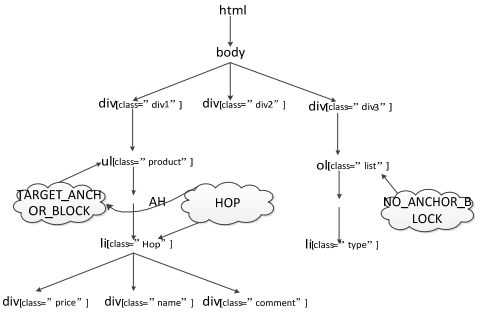

Figure 1. The Concept of Anchor-Hop-T Model

Anchor Point (AP). The base definition is equivalence to the concept of Anchor-Hop-I model, where the different is the manner of Anchor locating. For Anchor-Hop-I model, it adopts absolute path algorithm to locate anchor point using absolute path searching. This method neglects the relevance of tag, content and attribute of Hop. Therefore, the Anchor-Hop-T model is motivated to introduce new anchor locating algorithm in order to handle the failure of Anchor-Hop-I model.

Anchor Block (AB). In the domain of electronic products, the e-information displaied on Web page struct is usually in form of *table*, *ul* or *ol* tags. According to these characteristics, using XPath technology extracts the Web page content from *table*, *ul* or *ol* of sub-elements regardless of where the Web page struct is located. These sub-elements are called as anchor block, which consists of TARGET_ANCHOR_BLOCK and NO_ANCHOR_BLOCK. For the former, it involves the electronic products information list block that is what we want in information fusion system, including kinds of electronic products information, such as price, name, and comment. While for the latter, its procescing target is fouscing on non-target data block in Web page struct, inclduing non-electronic products information list.

Anchor Locating Library (ALL). In the anchor locating library, there are some extraction rules for kinds of commercial Website, which has all possbile paths for loacting anchor. The corresponding format is defined <Host,AnchorPath,FirstSuccTime,LastSuccTime,SuccNum,FailNum,Weight>, which will be explained in the Section 3. Note that the anchor locating library is constructed according to the order of Weight lable, where the more the weight value is the more reliability of anchor location is. In the information fusion system, it first to serach all anchor location paths which is mapped by the current rules saved in anchor location library. If it failed, the loacting algorithm, called as Anchor-T-Pos, will be used to autoamclly generate extraced paths. The new algorithm also will be updated as a template into the anchor location library, where the Weight calculation relates to FirstSuccTime, LastSuccTime, SuccNum and FailNum factor, which is detailed in section 3.3.

AH Relevance (AHR). AH relevance is the rate value between Anchor block and Hop point. Due to the Anchor block has two types, the TARGET_ANCHOR_BLOCK and NO_ANCHOR_BLOCK, how to distinguish and obtain these data of TARGET_ANCHOR_BLOCK and NO_ANCHOR_BLOCK needs to use AH relevance value. The greater AH has, the more the possibility of anchor block will be TARGET_ANCHOR_BLOCK. Based on this precondition, if AH relevance value is greater than the perdefined threshold value, the system will judge such anchor data as TARGET_ANCHOR_BLOCK. The correspoding algorithm is as follow. This computation is introduced in section 3.2.2.

III. The Implementation of Anchor-Hop-T Model

Anchor-Hop-T model is designed toward to handling Web page content and Web page structure of the Web electronic products Website, which has following two assumption conditions.

First, to the domain of electronic products, all information listed in each Web page is under the *table*, *ul* or *ol* tags. Based on these tags, our method of information fusion system uses XPath technology can conveniently extract anchor block.

Second, the Hop point can't be changed. The Anchor-Hop-T model doesn't improve the effectiveness of locating Hop point. But Hop point is still based on the absolute path of anchor point. Thus, the absolute path used for locating Hop point can't be changed. Even if it

© 2014 ACADEMY PUBLISHER

be changed, it can't impact the result of locating Hop point.

Under these assumptions, this section will first give the architecture of Anchor-Hop-T model, and then discuss locating anchor algorithm and the management of Anchor Locating Library.

A. The Architecture of Anchor-Hop-T model

The Anchor-Hop-T model considers the characteristic of Web page structure and handling the failure when using Anchor-Hop-I model. The architecture of Anchor-Hop-T model is as following Fig.2 shown. The anchor initial template liability saves the anchor locating library after successfully terminating system running, while it save the user's predefined extraction rules at the initial state of system running. The anchor locating update library module is used to the management of the initial template library, the persistent template library, and template library update. The locating module of Anchor-T-Pos handles and constructs the TARGET_ANCHOR_BLOCK of special Web page structure. Moreover, based on the Anchor-Hop-I model, it also extracts the related information of the locating path of current anchor in the anchor locating library.

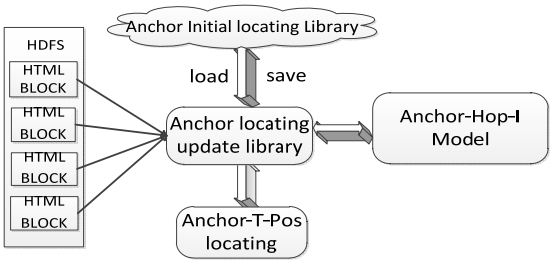

Figure 2. The workflow of Anchor-Hop-T Model

The workflow of Anchor-Hop-T model can be divided as follows: Step 1, execute the initialization, which is worked by anchor locating library that loads all anchor locating path from anchor initial locating library to memory anchor locating library. Step 2, caputre the original content of HTML from HDFS system, and further translate these contents into XML format in order to be modeled as DOM. Step 3, return the anchor locating path using DOM tree structure from anchor locating library. This information is pushed to the module of Anchor-Hop-I based information extractor. If it is success, anchor locating update module renews the relate anchor loacting path. After that, the process rocks back to step 2. If not, the next anchor loacting path will be picked up form anchor locating library for repeating this step. Step 4, if all corresponding anchor locating paths in anchor locating library has been traversed and the extractor can't correctlly gain any useful information, then Anchor-T-Pos loacting template generates a new anchor loacting path and anchor loacting library will revcevie this new path as its elements. Step 5, If the system terminates, the update module of anchor locating library persist the menory anchor locating library to the anchor initial locating library for the next time using after system restart.

B. Anchor Locator

How to locate anchor consists of obtaining anchor data and identifying TARGET_ANCHOR_BLOCK block. The latter is required to calculate AH value.

1) Obtaining Anchor Data

According to our investigation data, most of electronic products Website is in form of *table*, *ul* and *ol* tags in HTML display. Based on these tags, using XPath extracts paths possible containing multiple anchor block of target information. To traverse blocks, AH Relevance value is used that if the value is greater than the predefined threshold value the current anchor data block is considered as TARGET_ANCHOR_BLOCK. As Algorithm 1 shown, this algorithm is to obtain multi-data block list from HTML according to list tag of *ul*, *table*, and *ol*. After that, it is to obtain the subtitle target data block based on AH relevance Value.

Algorithm 1. Algorithm for obating TARGET_ANCHOR_BLOCK

Input: xmlContent(String): HTML is transferred to XML content
Output:List <Node> target Anchor data block
Procedure List<Node> getListAnchor(String xmlContent)
Step1: Accodirng to tags of ul,table and ol tags, using XPath to obtain corresponding multi-data block List<Node> ;
Step2: For(Node node in List<Node>);
Step3: Calculate AH relevance Value;
Step4: If AH is greater than threshold value Limit, than it is determined as TARGET_ANCHOR_BLOCK;
Step5: If AH is less than threshold value Limit, than it is determined as NO_ANCHOR_BLOCK;
step6: Return Anchor list;
End of **procedure**

2). AH Relevance Value and its Computing Algorithm

The AH relevance Value between Anchor and Hop coming from three parts of Hop, mainly relative path, attribute and content of Hop point, makes the Anchor-Hop-T model more effective to identifying TARGET_ANCHOR_BLOCK, where the parameters for each part are as follows:

Situation 1), The weight coefficient α of relative path starting from Hop point. As we know, the Hop path of Anchor-Hop-T model is constant. Thus, if the data can be extracted from Anchor data block based on Hop relative path, a strong possibility is TARGE ANCHOR BLOCK. To this point, the weighting coefficient is set as maximum.

Situation 2), The weight coefficient β of content. Due to the Anchor-Hop-T model faces on the Website of electronic products field, the list content of electronic products usually contains product price, category information. Moreover, it can't be influenced by the substantive changing of website structure. To this point, the weighting coefficient is set as maximum.

Situation 3), The weight coefficient γ of attribute. The attribute is possible be influenced by the changing of website structure. Therefore, its coefficient is lower compared to parameters α and β.

According to the introduction of influence coefficient in above, we can get following weight computation formula.

$$\psi = \alpha P+\beta C+\gamma A \qquad (1)$$

© 2014 ACADEMY PUBLISHER

where P is the goodness of fit for the relative path of Hop point, C is the goodness of fit for the content, and A is the goodness of fit for attribute. Their relation is defined as $\alpha \in [0,1]$, $\beta \in [0,1]$, $\gamma \in [0,1]$ and $\psi \in [0,1]$, where $\alpha + \beta + \gamma = 1$ and $\alpha > \beta > \gamma$.

The goodness of fit C includes kinds of content type, such as price and produce category. Thus, it needs formula to compute C. We mainly use following formula that matched number of content divides the total number of content types

$$C = \text{MatchtedContents}/\text{TotalContentsTypes} \quad (2)$$

Similarly, the goodness of fit A is calculated as follow, mainly involving matched number of attributes divides the total number of attributes.

$$A = \text{MatchtedAttributes}/\text{TotalAttributes} \quad (3)$$

Anchor-Hop-T model needs add tag in order to enchance the accuracy of extraction, using weight matching calculation of tag, attributes and content. When the value of weight is greater than some predefined threshold, it is considered as a success matching point. As the Algorithm 2 shown, the process works as follows:

Algorithm 2. Algorithm for obating Anchor block and weight value of relevance value Hop

Input: the string *curr* of Hop tag, attribute, and content in extraction rule. And node(Node): waited checking Anchor block. The format of *curr* is //tr[@class="class",contains("price")].);
Output: Similarity weights
Procedure float getSimilarity(String curr,Node node)
Step 1: Check whether the relative path of matching Hop point exits Anchor block, which is marked as *pathParam*.;
Step 2: If *curr* has content function, check whether the content is matched with node content, which is marked as *contentparam*;
Step3: If curr has attribute configuration, take out all attribute of original Hop point configuration. If not, skip Step6;
Step 4: Compare the all attribute in *attrList* with all attribute in sub-Node node;
Step 5: The attribute value is figured as the number of matched attribute divides the total number of attributes, which is marked as *attrparam*;
Step 6: According to the path, content and attribute for determining coefficient, weight computation formula is used to compute weight value;
Step7: Return weight value;
End of **Procedure**

C. Anchor Locator Management

The extraction rule of Anchor is defined as 7 tuple that <Host,AnchorPath,FirstSuccTime,LastSuccTime,SuccNum,FailNum,Weight>,where,

Host: is domain address of Website. For example, TaoBao Host is www.taobao.com, 360Buy Host is www.360buy.com. Host is mainly used to identify the corresponding relative path when using anchor to locating.

AnchorPath: is the absolute path of locating anchor. It is used as Anchor-Hop-I model's absolute path for extracting produce information. When a failure occurs in process of extracting absolute path for AnchorPath of Host, Anchor-T-Pos locater is used to generate new Anchor locating path.

FirstSuccTime: is to record the time for recording continuously extracting corresponding AnchorPath at first success situation. It combining with LastSuccTime can impact the current Weight value of AnchorPath.

LastSuccTime: is to record the time for continuously extracting corresponding AnchorPath at last success situation.

SuccNum: is total number of success extracted Webpages according to the current AnchorPath. Thus, considering the LastSuccTime factor, the more total number is, the more effectiveness of AnchorPath has. If it has high success probability in previous extraction and low failure probability in current extraction, it states that the AnchorPath method used in extraction lacks effectiveness, which calls for overall consideration.

FailNum, is total number of failure extracted Webpages according to the current AnchorPath. Comparing to the SuccNum, considering the LastSuccTime factor, the more total number is, the less effectiveness of AnchorPath has. If it has high failure probability in previous extraction and low success probability in current extraction, it states that the AnchorPath method used in extraction lacks effectiveness, which also calls for overall consideration.

Weight, is the weight value of current AnchorPath. It has more effectiveness of AnchorPath when the weight value is larger. In contrary, it has less effectiveness of AnchorPath. But, each AnchorPath exsiting in Anchor locating library has certain validity. Otherwise, when the weight value is less than certain threshold value, it needs to be removed out from Anchor locating library.

Weigh value consists of parameters <FirstSuccTime, LastSuccTime, SuccNum, FailNum>. The Weigh value is calculated by following situations:

The first situation is that Result = LastSuccTime + Threshold – CurrentTime, where Threshold is threshold value and CurrentTime is the current value. If the result value is less than 0, it means that AnchorPath is current failure. Coefficient α is set as ($\alpha<0$). The smaller Result is, the longer Anchor will have long failure time. On the other side, the weight impacted by the increased number of FailNum will be reduced. The weight value calculation formula is as follow:

$$Weight = Weight + \alpha\,(CurrentTime - ThreadShold - LastSucctime) * FailNum / (FailNum + SuccNum) \quad (4)$$

The second situation is that Result1= LastSuccTime + Threshold – CurrentTime,where Threshold is threshold value and CurrentTime is the current value. And Result2= FirstSuccTime+ Threshold – CurrentTime. If Result1 is greater than 0 and Result2 is less than 0, it means that the Anchor is effective for a long time. Coefficient β is set as ($\beta>0$). The more Result2 is, the more effective Anchor will get. Additionally, the weight impacted by the increased number of SuccNum will be added. The weight value calculation formula is as follow:

$$Weight = Weight + \beta\,(ThreadShold + FirstSucctime - CurrentTime) * SuccNum / (FailNum + SuccNum) \quad (5)$$

The third situation is that Result1= LastSuccTime + Threshold – CurrentTime, where Threshold is threshold value and CurrentTime is the current value. And Result2= FirstSuccTime+ Threshold – CurrentTime. If Result1 is greater than 0 and Result2 is greater than 0, it means that the Anchor fails in previous time but can work in current environment. This may impacted by the type of

© 2014 ACADEMY PUBLISHER

Webpage. Coefficient λ is set as λ (β>λ>0). At this time, Result2 is decrease, which means that AanchorPath impacted by type of Webpage but it tends to stable until turn to situation 2). Additionally, the weight impacted by the increased number of SuccNum will be added. The weight value calculation formula is as follow:

$$Weight = Weight + \lambda(1+1/(ThreadShold + FirstSucctime - CurrentTime)) * SuccNum/(FailNum + SuccNum) \quad (6)$$

According to above compution method, we can get following update algorithm for Anchor locating library.

Algorithm 3. Algorithm for updating Anchor locating library.

Input: host(Host) is address of Website. path(String) is the current Anchor locating path. result(boolean) returns the result for recording whether the extraction is success or failure.
Ouput: update the Anchor locating path
Procedure void updateAnchor(Sting path, boolean result)
Step 1: Obtain Anchor locating library Map<Host,List<WeightModel>> map;
Step 2: According host, obtain List<WeightModel>list of corresponding host from map;
Step 3: According path, obtain corresponding (WeightModel) model from list;
Step 4: If result is true, update SuccNum of model, and compute Weight. After turn to Step 6;
Step 5: If result is false, update FailNum of model, and compute Weight;
Step 6: According Weight, reorder list;
Step 7: Update map, and finish the process of Anchor locating library;
End of **Procedure**

VI. Experiment and Result Analysis

The experiment is designed and ran under Linux Ubuntu 11.04, and our prototype is developed by Eclipse3.5 IDE with JDK6.0. CUP is i5-2400 3.10GH, memory is 4GB. In our experiments, we focus on E-commerce Website taobao, Pacific and 360buy as template website in information extraction. Each Website has huge Electronic product data, where the product list information of Webpage can be used as information source for template Webpage. The main ideal is to use Web crawlers to capture Webpage information and take out contrast experiment for DOM-based Anchor-Hop model, Anchor-Hop-I model and Anchor-Hop-T model for point of view at recall ratio, precision ratio and time-consuming during extraction process.

A. Information Evaluation Criterion

The evaluation criterion we used in extraction experiments includes recall ratio, precision ratio. Their formal defiend as follows:

Recall Ratio= Extracted Correct Information Number / Total Correct Information Number (7)

Precision Ratio= Extracted Correct Information Number / Total Extracted Information Number (8)

where the number of Total extracted information is the number of extracted information from template Website in experiments. This information includes correct, and fault. Thus, the set is defined as {ALL_INFO}. The number of Total Correct Information is the number of correct e-produce from template Website in experiments. The corresponding set is defined as {ALL_SHOULD_INFO}. The number of Extracted Correct Information is the number of correct information from template Website in experiments. The Recall Ratio and Precision Ratio is between 0 and 1. When the Precision Ratio is high and the number of extracted information is low, it is possible omit to extract some information, which results in the reduction of recall ratio.

During comparing performance under different extraction models, it requires to consider following factors. In order to integrate them and compare each other, the weighted collection average value F is used for Recall Ratio and Precision Ratio.

$$F = (\beta^2 + 1)PR/(\beta^2 P + R) \quad (9)$$

In this formula, P is Recall Ratio, R is Precision Ratio and β is the impact factor in F when investigating P and R. When β=0, F is related with F but unrelated with R. When β tends to infinity, F is only related to R but unrelated with P. Without loss of generality, β is set as β=0 to evaluate the effectiveness of information extraction system. Except above three indicators, we also discuss to compare extraction effectiveness focusing on time-consuming of extraction.

B. Experiment Processing

Our experiments adopt the search entrance of taobao, Pacific and 360buy as sample for statistics extraction. In these Website, we input keyword “笔记本” in Chinese, where taobao display 100 relate Webpage with 40 entry in each Webpage, Pacific display 165 relate Webpage with 25 entry in each Webpage and 360buy display 155 relate Webpage with 25 entry in each Webpage.

1) Information extracted from Taobao Website

In Taobao Website, the limit entrance http://s.taobao.com/search?q=%B1%CA%BC%C7%B1%BE can be used as the extract entrance template. According to the structure of taobao Webpage, we can get following extraction rules.

1) The rule of Anchor-Hop model can use attribute-based locating method. The Xpath search of Anchor is //ul[@class=”list-view”] and the Xpath search of Hop is../li[@class=”list-item”].

2) In DOM-based dynamic Anchor-Hop-I model, Anchor point is located by absolute path. The Xpath search of Anchor is /html/body/div[@class=”mall”]/ul[@class=”list-view”] and the Xpath search of Hop is ../li[@class=”list-item”].

3) The path of Anchor locating library is initialed by DOM-based Anchor-Hop-T model. Due to the possible failure of path, the Anchor path can be generated by Anchor locating library. The Xpath search of initial Anchor point in Anchor locating library is /html/body/div[@class=”mall”]/ul[@class=”list-view”] and the Xpath search of Hop is ../li[@class=”list-item”].

2) Information extracted from Pacific Website

In Pacific Website, the limit entrance http://product.pconline.com.cn/notebook can be used as the extract entrance template. According to the structure of Pacific Webpage, we can get following extraction rules.

© 2014 ACADEMY PUBLISHER

1) The rule of Anchor-Hop model can use attribute-based locating method. The Xpath search of Anchor is //ul[@class=" product-list"] and the Xpath search of Hop is../li.

2) In DOM-based dynamic Anchor-Hop-I model, Anchor point is located by absolute path. The Xpath search of Anchor is /html/body/div[@class="content"]/ul[@class=" product-list"] and the Xpath search of Hop is ../li.

3) The path of Anchor locating library is initialed by DOM-based Anchor-Hop-T model. Due to the possible failure of path, the Anchor path can be generated by Anchor locating library. The Xpath search of initial Anchor point in Anchor locating library is /html/body/div[@class="content"]/ul[@class=" product-list"] and the Xpath search of Hop is ../li.

3)Information extracted from 360Buy Website

In Pacific Website, the limit entrance http://search.360buy.com/Search?keyword=%B1%CA% BC % C7% B1%BE can be used as the extract entrance template. According to the structure of Pacific Webpage, we can get following extraction rules.

1) The rule of Anchor-Hop model can use attribute-based locating method. The Xpath search of Anchor is //ul[@class="list-h clearfix"] and the Xpath search of Hop is../li[@sku="359578"].

2) In DOM-based dynamic Anchor-Hop-I model, Anchor point is located by absolute path. The Xpath search of Anchor is /html/body/div[@class="m"]/ul[@class=" list-h clearfix"] and the Xpath search of Hop is ../li[@sku="359578"].

3) The path of Anchor locating library is initialed by DOM-based Anchor-Hop-T model. Due to the possible failure of path, the Anchor path can be generated by Anchor locating library. The Xpath search of initial Anchor point in Anchor locating library is /html/body/div[@class= " m"]/ul[@class=" list-h clearfix"] and the Xpath search of Hop is ../li[@sku="359578"].

C. Result Analysis

According to the above data, we can get following conclusions:

1) In these three models, Anchor-Hop model has low precision ratio, which occurs failures in Taobao and 360Buy. From analyzing the element of Taobao and 360Buy, we can see that in Taobao Website the tag *li* of attribute class="list-item" is 4127. And in 360buy Website the tag *li* of attribute sku="359578" is 4368.

2) These three models have same Recall Ratio of 100%. Anchor-Hop model adopts global matching in Webpage, where Recall Ratio is high. Under unchanged structure Website, Anchor-Hop-I model can accurately locate Anchor point and Hop point, which has high matching efficiency. Because Anchor-Hop-T model is an extension of Anchor-Hop-I model, they all has high recall ratio.

3) The time-consuming for models is that Anchor-Hop time is greater than Anchor-Hop-T. Due to the global matching in Webpage, the efficiency is low. Anchor-Hop-I model has the best efficiency due to its absolute path matching method. Anchor-Hop-T model refers to Anchor locating library, which results in low efficiency compared to Anchor-Hop model.

TABLE1.
THE RESULTS OF EXPERIMENT

Method / Website	Number	Correct	Precision Ratio	Recall Ratio	Time	F
Anchor-Hop / Taobao	4127	4000	96.92%	100%	48053ms	0.9843
Anchor-Hop-I / Taobao	4000	4000	100%	100%	28917ms	1
Anchor-Hop-T / Taobao	4000	4000	100%	100%	30681ms	1
Anchor-Hop / Pacific	4239	4107	96.88%	100%	5044ms	0.9841
Anchor-Hop-I / Pacific	4107	4107	100%	100%	30721ms	1
Anchor-Hop-T / Pacific	4107	4107	100%	100%	32047ms	1
Anchor-Hop / 360Buy	4368	3875	88.71%	100%	46719ms	0.9401
Anchor-Hop-I / 360Buy	3875	3875	100%	100%	25349ms	1
Anchor-Hop-T / 360Buy	3875	3875	100%	100%	28349ms	1

V. CONCLUSIONS

In order to implement the information fusion of electronic products, this paper proposes a DOM-based Anchor-Hop-T information extraction model for Web application. It combines Anchor block and AH relevance value to correctly generate Anchor locating path for e-produce information extraction. Through experiments among Taobao, Pacific and 360Buy Websites, our proposed approach is demonstrated to be applicable in the information extraction. Considering that the general search engine returns a large amount of information for user without any processing, the most current retrieval efficiency is low. In the future, based on the extracted information, we will continue to develop the distributed information retrieval system for retrieving Web electronics products information based on Hadoop and Lucene. As our next research plan, the Hadoop is considered as the base supporting tool for the distributed

© 2014 ACADEMY PUBLISHER

retrieval and Lucene is used as the search engine using the inverted index technique.

ACKNOWLEDGMENT

The Project Supported by Zhejiang Provincial Natural Science Foundation of China under grant No.LY12F02029, National Technology Support Program under grant No.2011BAH16B04, and Zhejiang Provincial Scientific Research Project.

REFERENCES

[1] Freitag D ,McCallum A, “Information extraction with HMM structures learned by stochastic optimization", In Proceedings of the Eighteenth Conference on Artificial Intelligence, pp.584-589, 2002.
[2] Souyma Ray ,Mark Craven, “Representing sentence structure in hidden Markov models forinformation extraction”, In Proceedings of the Seventeenth International Joint Conference OnArtificial Intelligence, pp.1273 -1279, 2001.
[3] Steve Lawrence and C. Lee Giles and Kurt Bollacker, “Digital libraries and autonomous citation indexing”, IEEE COMPUTER,Vol.32,No.6,pp.67-71,1999.
[4] Qian Mo and Yi-hong Chen. Ontology-Based Web Information Extraction. ICCIP 2012, Part I, CCIS 288, pp.118-126, 2012.
[5] Xue Jian, "Research on Web Information Extraction Based on the Improved Maximum Entropy Algorithm", AISS, Vol. 4,No.13, pp.85 -91, 2012
[6] Rong LI, Chun-qin PEI, Jia-heng ZHENG. Web Information Extraction based on HybridConditional Model. 2010 Second International Workshop on Education Technology and Computer Science, pp.137-141, 2010.
[7] Soderland S. Learning Information Extraction Rules for Semi. structured and Free Text.Machine Learning.1999.
[8] Eirinaios Michelakis, Rajasekar Krishnamurthy, Peter J. Haas, and Shivakumar Vaithyanathan. Uncertainty management in rule-based information extraction systems. In Proceedings of the 2009 ACM International Conference on Management of data (SIGMOD '09), pp.101-114, 2009.
[9] XML Path Language (XPath) 2.0. http://www.w3.org/TR/xpath20/.

Yuanyuan Zhang M.S. degree, lecturer at College of Information Technology, Zhejiang Chinese Medical University. Her research includes Medical Middleware and Web Information Mining.

© 2014 ACADEMY PUBLISHER