http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 13 (1): 69-77, 2014
ISSN 1812-5638 / DOI: 10.3923/itj.2014.69.77

Research on Extraction Methods of Web Page's Document Logical Structure

[1,2]Wei Wang, [2]Wei Wei, [1]Qinghua Zheng, [2]Jie Hu, [1]Yingying Chen and [3]Bin Zhou
[1]School of Electronic and Information Engineering,
Xi'an Jiaotong University, Xi'an 710049, China
[2]School of Computer Science and Engineering,
Xi'an University of Technology, Xi'an 710048, China
[3]College of Science, Xi'an University of Science and Technology,
Xi'an, 710054, China

Abstract: Based on the analysis of characteristics of web page data set and difficulties of document logical structure extraction task, the method of document logical structure extraction of web page is proposed, moreover, four key technologies are proposed in order to extract document logical structure. Finally, the study download and process a number of web pages from Baidu baike and general sites related to two courses of computer science i.e., operating system and computer network. Evaluation on web pages of Baidu baike shows that the average error rate is 12.8 and 6.6% on operating system and computer network courses respectively and the average rate of general web pages on operating system and computer network is 30 and 22.6%, respectively. The experimental results validate the effectiveness of the method proposed in this study.

Key words: Document logical structure, web information extraction, minimum semantic logical block, optimal sequence solving

INTRODUCTION

Information extraction technology research first began in the mid-1960s, have been attracted much attention because of its broad application prospects. Information extraction (Information Extraction) technology refers to the application directly extract information from natural language text, the form of a structured description for information queries, text deep excavation, automatic answer questions, access tools to provide people with a strong message (Wei *et al*., 2011). Its main function is to extract structured information from unstructured or semi-structured documents to inquiries and next. Its early study a variety of forms, including news reports, medical records of patients, Word or PPT format, such as educational resources, scientific and technical literature and company reports. And general Web information extraction object, the Web page of the document logical structure throughout the entire Web page (Mao *et al*., 2003) facing some new challenges, including.

General web information extraction object is usually specific information, a single line of text or more consecutive lines of text, the document logical structure usually contains more than one line of text and span, multi-level features. The multi-level document logical structure contains multi-level sub-headings, such as 1.1, 1.1.1, its feature extraction and description in the form of relatively difficult (Ashish and Knoblock, 1997). Based on document logical structure of the multilayered nature, it needs to draw on existing Web information extraction technology, so that present a new idea.

MATERIALS AND METHODS

Data preprocessing: The data set is the validation of models and methods. In order to reflect the diversity of the data set, the study from Baidu Encyclopedia, general site to download Web pages with a certain scale computer field. Prior to describe the data pretreatment process (EL-Shayeb *et al*., 2009), the study first gives the definition of the candidate sub-title.

Definition 1: Candidate sub-title refers to a Web page topic relatively independent natural paragraphs denoted by h.

Web page data pre-processing can be divided into a total of three stages: Noise information filtering, candidate sub-title generation and data labels.

Corresponding Author: Wei Wang, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049, People's Republic of China

Based on the above analysis, the study can see the Web page pre-processing tasks can be formalized defined as follows:

Preprocess: *OrgHtml*→*H*
In formula:
Preprocess——*pretreatment process*;
OrgHtml —— The original Web page, can be expressed as: *OrgHtml* = (*MainInfo*, *AdInfo*, *NaviInfo*, *CopyrightInfo*);
H——the candidate sub-title set.
Noise information in data preprocessing filter stages can be formalized as follows:
NoiseFilter:OrgHtml→
Which *CleanHtml* = (*MainInfo*),represents a clean Web page。
Data pre-candidate sub-title generation phase can be formalized as follows:
CandidateSubTitleExtrator: CleanHtml→*H*

With the level of sub-title recognition title Solution

Feature extraction

Format features: The format is characterized by the characteristics of the candidate sub-header format, candidate sub-title font size, font color, bold, whether it is italic and so on. For example, if a candidate sub-title of the Web page is bold, or font color is black, the candidate sub-title is likely to be sub-headings (Kurland and Lee, 2005). All the characteristics of candidate sub-title set for all the Web pages there are some limitations of this global characteristics. Not unified format between pages, such as the author of page A H4 to represent ordinary text, page B of H2 text. H2 in A are used to modify the sub-title but in B just plain text font size. Therefore, the different page font size for the type of Web page showing the different severity levels is not comparable. The global features mentioned in the above table the normalization processing, to generate a local characteristic, that is, to consider the various features in a single Web page internal calculation. For example, this feature of the font size, the first traverse of the font size of a single Web page (Zhang *et al*., 2003), the statistics of the most of the candidate sub-title font size as the font size, according to the difference of the actual font size font size of each of the candidate sub-title, local characteristics.

Context characteristics: Context characterized candidate subtitle context information, such as change of the previous or next candidate sub-title font size, font name change, color change (Witten and Frank, 2005). This format change reflects this degree of importance of the candidate sub-headings.

Definition 2: Let terms constitute a collection of candidate sub-heading (NATURAL paragraph) appears on average word in the candidate sub-heading (NATURAL paragraph), such as the Eq. 1:

Table 1: Candidate linguistic characteristics of the sub-title

Feature item	Meaning of the feature items	Type
TitleCue	Candidate sub-title cue word	Binary
NonTitleCue	Candidate sub-title before the prompt word the non-subtyped title	Binary
HasPunc	Candidate sub-heading at the end of the end punctuation	Binary
Words	he No. of sub-word	Real
Terms	No. of terms	Real
NumOrLetter	Whether the candidate sub-heading are No. or letters	Binary
AverTFBe	In terms of the sub-title of the candidate appears in the sub-header on a candidate on average word f (T_i, h_{i-1})	Real
AverTFAf	In terms of the sub-title of the candidate, the next candidate for the sub-title on average word	Real

$$f(T_i, p_j) = \frac{1}{|T_i|}\sum_{t_k \in T_i} tf(t_k, p_j) \quad (1)$$

where, t_k-T. In the first term, tf (t_k, p_j). In candidate sub-headings (Natural paragraph) word frequency.

Given below detailed linguistic characteristics, as shown in Table 1.

Level of characteristics: Level features is able to reflect the characteristics of the sub-heading level. Linguistic features mentioned above only describes the candidate sub-title is the characteristics of the sub-title but cannot explain the first few levels of sub-headings. Level characteristics from the characteristics of the primary, secondary and tertiary sub-title to said candidate sub-title.

Sub-heading level the optimal sequence solving: On the basis of the identification of the band level of the sub-title, the title of this section pair category identification easy-to-digest followed by three correction algorithm, in order to further improve the recognition performance of the sub-heading level (Joachims, 2002) reduce the error of the subsequent structured document logical structure tree.

The sub-title of the Web page set as a whole of the input model of learning in the entire sub-title set constituted on the basis of the sequence, thereby obtaining a sequence of prediction model. It can be that the optimal sequence of sub-heading level solution is to establish a conditional probability distribution between the output sequences of tokens from the input sample sequence to the process, such as the representation of this distribution. Wherein the input sample sequence is the sequence of the sub-title of the Web page set, the output token sequences correspond to the level of the sub-title set. When given a new input sequence (Dumais, 1998), the algorithm searches in the solution

space of the whole of the output sequence and the probability of one of the largest as a final output sequence, such as the Eq. 2.

$$y_{i[1:T]} = \underset{Y_{[1:T]} \in \mathcal{Y}}{\arg\max} P(Y_{i[1:T]} \mid x_{i[1:T]}) \quad (2)$$

If a Web page is the number of the sub-title for t, the value of each sub-title C′ = *{FirstLevel, SecondLevel, ThirdLevel}*, the entire solution space of the output sequence of size 3^t. When the number of the sub-title of the Web page increases, the size of the corresponding solution space is growing exponentially. Therefore, the following three easy-to-digests followed by discussion of how to reduce the solution space, excluding a large number of output sequence could not exist and with a level of sub-title the results of the identification phase to find the optimal output sequence.

Logical constraint-based conflict detection and correction: With the level of sub-headings to identify candidate sub-title set to obey the premise of independent and identically distributed, the presence of only consider the characteristics of the candidate sub-title, ignored before and after the candidate sub-title which makes the identified sub-heading level the apparent conflict. For example, identify the front and rear sub-title there is a substantial level across the first sub-title of the Web page is not a sub-heading and so on. Therefore, the initial identification of the sub-title-level constraints amendment to the conflict that appears, on the one hand, in order to resolve the apparent conflict, the sub-title set more reasonable level (Manevitz and Yousef, 2001), on the other hand, in order to build document logical structure correctly establish the connection between nodes in the tree.

The level of sub-headings need to be both logical constraints, on the one hand from the Web page level constraints the level of the sub-title; party is to constrain each sub-heading level from the front and rear sub-heading level combination. Therefore, the sub-heading level logical constraints are the following:

Constraint 1: The first child of each Web page first title to one level, this is:

$$\forall i \in N,\ y_{i1} = \text{FirstLevel},\ y_{i1} \in y_{i[1:T]}$$

Constraint 2: If the current title is the K level is close to the title, the title back level cannot be greater than or equal to k+2, this is:

$$\forall i \in N,\ \forall j \in T,\ y_{ij} = k \rightarrow y_{ij+1} < k+2$$

Table 2: Logic of conflict detection and correction algorithm based on constraint

Algorithm: Based on logical constraints conflict detection and correction algorithm
Input:
SubTitleSet: Said single post Web pages have been identified with the level of the sub-title collection
Output:
Definition:
$SubTitle_i$: That the Web page set of conflict detection and correct the sub-title
Definition:
$SubTitle_i$: Which means that the title of the i sub-band level
The steps of the algorithm:
Do{
For Each sub-title $subTitle_i \in SubTitleSet$
(I = 1,2, ..., n);
If Current sub-title $subTitle_i$ Violation of logical constraints that conflict
Select with probability times the level in the level sub-title recognition stage as its new level
}while(Sub-title sequence conflict)
The end of the algorithm

The basic idea of logic constraints conflict detection and correction algorithm based on the logical constraints is to detect conflicts have subheadings in the sequence and the corresponding strategies to modify the conflict, the detailed algorithm description as shown in Table 2.

Conflict detection and correction algorithm based on logical constraints can take advantage of the front with the level of sub-headings recognition results of the identification phase, narrow the solution space size, simplify processing steps while the development of constraint is also relatively easy. However, the limitations of the point of the algorithm: Have different characteristics due to the different types of Web pages, it is therefore necessary to develop different constraints on different types of pages, while exhaustive constraint more difficult and often not reasonable (Cauwenberghs and Poggio, 2001). If you can uniform treatment of different types of Web pages, select a common solution model will better achieve the pairs title optimal sequence solving.

K-order optimal sequence solution: A description of the section with a general solution to a model to solve the optimal sequence of sub-headings. This model need to address two aspects, on the one hand, take into account both the differences between the different Web pages but also to consider the correlation between the same Web page within the sub-title; the other hand, is to narrow the solution space size (Weston *et al*., 2001). The following will start the detailed description.

This section describes the optimal sequence of sub-heading level to solve the problem. To solve the linear sequence of the sub-title of a Web page set the converted to solve the problem: a linear sequence of a known input in the solution space of the corresponding sequence of tokens, to find the probability of a sequence

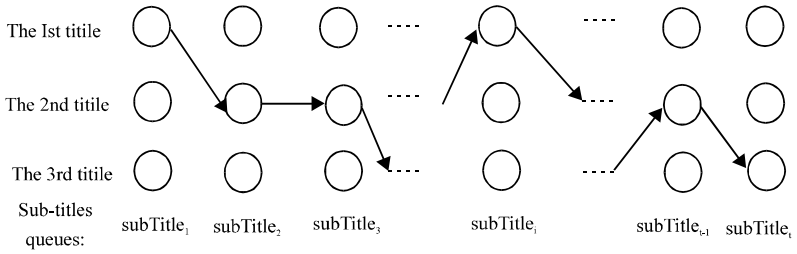

Fig. 1: Schematic diagram of the sequence of all the solution space neutron title level to solve

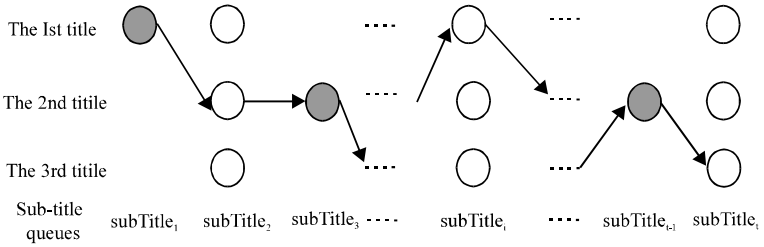

Fig. 2: Based on a schematic diagram of the sub-title of the threshold level sequence solving

of tokens and output. Figure 1 shows the schematic diagram of solving level in all solution space Neutron title sequence.

Figure 1, enter the sub-title sequence is a sequence of sub-title set for a Web page, a total of t sub-title, the output corresponding to the target category is sub-headings, arrows indicate the change process of the sub-title sequence of the target class. Can be observed from Fig. 1, with the sub-title sequence length increases, the number of the output sequence may increase in an exponential scale. Solving the optimal sequence of the input sequence, the need to calculate the conditional probability of the current sub-title, if you consider this sub-title where the entire sequence of information, will inevitably increase the computational complexity but if only the calculation of the sub-title of this sub-title-context information is not enough, lose some critical information, resulting in errors of judgment. Therefore, based on the above considerations, the K-order characteristics to consider the current sub-title of context information, which uses the current sub-title K sub-title of the sub-title target category judgment (Amari and Wu, 1999). Here is the K-order optimal sequence formal description of the method for solving.

In K order optimal sequence solving, the conditional probability of the sequence can be expressed as:

$$p(Y_{i[1:T]} \mid x_{i[1:T]}) = \prod_{r=1}^{n} p(x_{ir} = y_{ir}) p(x_{ir} \mid x_{ir-1} = y_{ir-1}) p(x_{ir} \mid x_{ir-2} = y_{ir-2}) \ldots \prod_{r=1}^{n} p(x_{ir} \mid x_{ir-k} = y_{ir-k}) = \prod_{r=1}^{n} \prod_{s=1}^{k} p(x_{ir} = y_{ir}) p(x_{ir} \mid x_{ir-s} = y_{ir-s}) \quad (3)$$

Equation 3 in the value of K dimensional transition matrix, the starting node transition matrix on S and end node transition matrix E. Calculated conditional probability of all possible sequences in the solution space, the search for the maximum probability:

$$y^{*}_{i[1:T]} = \underset{Y_{[1:T]} \in y}{\operatorname{argmax}} P(Y_{i[1:T]} \mid x_{i[1:T]}) \quad (4)$$

Figure 2 shows the schematic diagram of a threshold based on the sequence of the sub-heading level to solve. The gray node the classification probability exceeds a set threshold node, sub-heading level can be determined, greatly reducing the number of unknown level of the sub-title, thereby reducing the size of the solution space.

Based on the above analysis, the the K order optimal sequence solving algorithm is described as shown in Table 3.

K-order algorithms for solving the optimal sequence is divided into two phases: Phase 1 initialize the corresponding parameters calculated by the training set three transition matrix; Phase 2 selection probability exceeds the threshold value and category as the label of the sub-title, reduce the unknown level the number of sub-header, followed by the calculation of the conditional probability of all the sequences in the solution space, the greatest probability of a sequence of search criteria will be returned.

The K-order optimal sequence algorithm not only uses the threshold to narrow down the size of the solution space but also uses the K-order characteristic of the

Table 3: K-order optimal sequence algorithm

Algorithm:	The optimal sequence algorithm order K
Input:	
SubTitleSet:	Expressed predict a single Web page has been identifiedin the level of subheadings collection;
HtmlTrainSet:	Expressed as a training set of pre-processing and labeling Web page;
Output:	
SubTitleSet´:	Said the Web page used to predict the optimal sequence of sub-heading level;
Defined:	
$SubTitle_i$:	Expressed the i subtitle with level;
Π:	Expressed K-dimensional transition matrix;
S:	Expressed starting node metastasis matrix;
E:	Expressed terminate node transition matrix;
l:	Expressed the category of the subtitle, l∈C´ = {FirstLevel, SecondLevel, ThirdLevel};
n:	Expressed the Web page in the number of unknown level of subheadings;
Algorithm steps:	
Stage 1:	
Step 1:	Manually set threshold ω, Initializes the transfer matrix Π, S and E;
Step 2:	Traverse the training set *HtmlTrainSet*,Calculation of transition matrix Π, S and E and its normalization processing;
Stage 2:	
Step 3:	Normalization n = 0;
Step 4:	To each child SubTitle in sub title of *SubTitleSet* If the max probability P p(subTitle$_i$ = l)>ω,Then, the category for the subtitle l; otherwise, n^{++};
Step 5:	The size of 3^n each sequence in the solution space: According to the formula (3-6) calculate the sequence of conditional probability;
Step 6:	Sequence of conditional probability maximum search solution space, to return it.
End of the algorithm	

current sub-headings and takes full account of the context characteristics. At the sanie time, the method contains several logical constraints conflict detection and correction method based on logical constraints.

For exanıple, constraint, the first sub-title of each Web page is a sub-title. Level sub-headings, so the matrix S for initial node metastasis calculated according to the training set because the training set of all the Web pages of the first sub-title, the probability level sub-title 1 while the other two sub-title probability are 0, then according to the Eq. 3, when r = 1 and y_{ir} = FirstLevel, p (x_{ir} = y_{ir}) = 0◦ If the title of the first child of a sequence is not a sub-title, the sequence of conditional probability is 0; the sequence cannot be the optimal solution. Therefore constraints reflected in the initial node transfer matrix S.

Such as constraints, 2, if the current sub-title is the k-level sub-title, the very next sub-heading level cannot be greater than or equal to k+2◦ K-dimensional transition matrix Π of the first dimension in steps of 1 (close to) the sub-title the transition probability between a matrix (Rish, 2001). Shows that the probability of a sub-heading to the conversion of three sub-headings in the first dimension of the transition matrix calculated according to the training set to 0. A possible sequence adjacent to the sub-title to convert from one to three, according to the Eq. 3-6, p (x_{ir-1}| = y_{ir-1}) = 0, hat the conditional probability of the sequence is 0, so the sequence cannot be the optimal solution. It is constrained reflected in the K-dimensional transition matrix Π.

Based on solved the sub-structure of the optimal sequence: On an organization, the sub-title set as a linear sequence of K-order characteristics of the sub-title as context information to consider its K sub-title on the sub-title linear. This method can better solve the optimal sequence of sub-heading level and to avoid some of the problems and shortcomings of conflict detection and correction method based on logic constraints. But concluded that: These sub-headings in the logic from the perspective of a document logical structure tree, based on the tree structure to organize, by its location in the document in the logical structure tree of the current level of the sub-title, including its parent node or its sibling. Ignored by the previous section and the characteristics of the Web page document logical structure tree, without considering the context information of the sub-headings in the tree. Therefore, in this section for the successes and shortcomings of the method of document logical structure tree, starting from the tree structure to seek the optimal solution of the sub-title sequence of a single Web page (Lewis, 1998). First, some related concepts.

Related concepts introduced

Definition 3: The smallest semantic logic block (Minimum Semantic Logical Block) $b_i = (t_i, P_i)$ (i = 1, ..., n) is composed of natural paragraphs set the Web page neutron title and follow up non-subtyped title set $P_i = \{p_{ij}|j = r, ..., k\}$. t_i is Not empty, P_i can be empty set, n for the number of the title of the Web page neutron, r behind t_i the first non-subtyped title paragraph number, k behind t_i is the paragraph numbers the last non subtype title.

It shows an exanıple of a Web page, the minimum semantic logic block. The smallest semantic logic block is an independent unit and this is because behind the sub-title the non-subtype title sets are often sub-title Subjects will be described. Minimum semantic logic block has a richer content than the sub-title, more expressive. This section uses the minimum semantic logic block as a node to represent document logical structure tree. The study shows the logical structure tree in Fig. 1. Web page corresponding to the new document. Each minimum semantic logic block of the subscript label indicates that it contains sub-title tag.

Minimum semantic logic semantic relationships between departments and regions, illustrates the intrinsic relationship between the minimum semantic logic blocks, such as the father-child relationship, sibling relationships and children and grandchildren-the ancestral relationship. For example, the minimum semantic logic block b2 is b1 subordinates, b1 and b2 are semantically father-child relationship (McCallum and Kamal, 1998). b3 and b4 is the same level of minimum semantic logic blocks, b3 and b4 are brothers. B12 level significantly precedence over b11, b11 and b12 descendants-ancestral relationship.

Definition 4: Minimum semantic logic block collection of type of semantic relations BRT can be expressed in the Eq. 5:

$$BRT = \{father\text{-}children, brother, children\text{-}ancestors\} \quad (5)$$

Set ($<b_i, b_j>$, r), i<j minimum semantic logic block semantic relations, among $<b_i, b_j> \in B \times B$, B is a minimum semantic logic block set: r∈BRT:

- r = father-children which means that the minimum semantic logic block b_i is the parent node in the document logical structure tree upper b_j is a child node, located in the lower of the document logical structure tree
- r = brother which means that the minimum semantic logic block b_i and b_j are the nodes at the same level, that the two brothers
- r = descendants-ancestors which means that the minimum semantic logic block b_i in the Web page is located in the front of the b_j in the document logical structure tree in the document logical structure tree b_i lower b_j is located in the upper layer of the document logical structure tree

Basic idea of the algorithm: This section departure from the sub-structure of the document logical structure tree, the optimal sequence of the K-order method for solving the sub-structural features combine to seek the optimal sequence in the solution space (Quinlan, 1986). Take into account not only the current sub-title, their linear sequence on the K-order features and also consider the sub-structure of their document logical structure tree.

The basic idea of the algorithm based on the sub-structure of the optimal sequence is: According to the semantic relationship between the minimum semantic logic block and the immediately preceding minimum semantic logic block, adjust the conditional probability of possible sequences (Quinlan, 1996), in order to find the maximum probability of the sequence as the optimal solution.

In the optimal sequence solving improved, according to the minimum semantic logic block on the semantic relationships, the conditional probability of the sequence can be expressed as the Eq. 6:

$$p(Y_{i[1:T]} \mid b_{i[1:T]}) = \prod_{r=1}^{n} p(r(b_{ir-1}, b_{ir}))p(b_{ir} = y_{ir})p(b_{ir} \mid b_{ir-1} = y_{ir-1})... \prod_{r=1}^{n} p(b_{ir} \mid b_{ir-k} = y_{ir-k}) = \prod_{r=1}^{n}\prod_{s=1}^{k} p(r(b_{ir-1}, b_{ir}))p(b_{ir} = y_{ir})p(b_{ir} \mid b_{ir-s} = y_{ir-s}) \quad (6)$$

Where:

- $\mathbf{b_{i[1:T]}}$**:** The set of the smallest semantic logic block i of Web page
- $\mathbf{b_{ir}}$**:** The r of i Web page smallest semantic logic block
- **r** $(\mathbf{b_{ir-1}, b_{ir}}) \in$**BRT:** The semantic relationship between the b_{ir-1} the minimum semantic logical block b_{ir-1} and b_{ir} type

Priori probability p (b_{ir} = y_{ir}) and conditional probability p ($b_{ir}|b_{ir-s}$ = y_{ir-s}) of the calculation in the same the K order optimal sequence solving (Bille, 2005). Determine the type of semantic relationships that exist between them according to the level of the b_{ir-1} and b_{ir} type, is the probability of the three relationships according to the minimal semantic logic block identification of semantic relations relationship the algorithm calculated b_{ir-1} and b_{ir}. Then, p ($r(b_{ir-1}, b_{ir})$) of the probability is when r = type (Wei *et al*., 2010a).

How to identify the minimum semantic logic blocks on the type of semantic relations? The minimum semantic logic block relationship recognition as a multi-class classification problems in machine learning (Salton *et al*., 1975), the upcoming minimum semantic logic block as a classification instance, classification category the BRT = {father-children, brothers, children-ancestors}. Given under Chapter II machine learning classification of the working mechanism of the whole relationship identification process can be divided into the feature vector generating the discriminate model generation, model to predict the three processes. The discriminate model generation refers to the use of the process to determine the classification of the feature vector generated by the training set. The model predictions are learned classifier is used to identify the type of semantic relations unknown minimum semantic logic block. The following describes the feature extraction or feature vector generation (Wei *et al*., 2012a).

Minimum semantic logic blocks appear as text, cannot be the computer identification, so the vector space model, the minimum semantic logic block feature representation. Selected through the draw with a level of sub-title identification characteristics, the study selected

characteristics of the format characteristics, level features and linguistic features to represent the minimum semantic logic block. (<b_i, b_j>, r), i<j.

Format characteristics: Because of the comparability between the two smallest semantic logic block is mainly embodied in natural between paragraphs, based on such considerations, the formatting characteristics is mainly reflected in the sub-title, non-child format comparison between the title.

Document logical structure tree algorithm: Identified when the level of the sub-title of candidate Web page, followed by the build document logical structure. The sub-title of the function of this part is based on the level of the band has been identified as an input, to construct a corresponding Web page document logical structure tree (Wei *et al*., 2010b).

Document logical structure tree is a directed tree, non-root node in the tree representation of chapters, sections, subsections, sub-headings. From close to the front of the subTitle$_i$ node scan when new nodes subTitle$_i$ document logical structure tree. First determine the *subTitle$_i$* level, when it is a *subTitle$_i$*, the document logical structure tree, become a child of the root node of the root, at the same time to do the appropriate parameter adjustment (root children plus 1 that is *root.cn*++, root a child node of the root of cn subTitle$_i$ root.childcn. = *subTitle$_i$*, *subTitle$_i$* parent node is the root of *subTitle$_i$*. *parent* = *root*, the level is *subTitle çn* is *root.level*=*cn*), when it two sub-headings to find close to its previous node subTitle$_{i-1}$ judgment subTitle $_{i-1}$ is a sub-title, until you find a *subTitle$_i$*, Add *subTitle$_i$* to become a sub-heading tree children node; when subTitle$_i$ three *subTitle$_i$*, found along it to the upper layer of the tree to find until two sub-headings and added to the tree as child nodes of the two sub-headings (Wei and Jun, 2012). Document logical structure tree construction algorithm is described as shown in Table 4.

Table 4: Documant logical structure tree condtruction algorithm

Algorithm: Documant logical structure tree building algorithm
Input;
SubTitleSet: Said a single Web page has been identified with subheadings collection level;
Output:
DLST: Said the Web page document logical structure of the tree;
Definition:
subTitle$_i$: Said the ith a subtitle with level;
root: Said DLST root node;
Algorithm steps:
Step 1: Traverse the entire candidate subtitle, will subheadings in the paragraph on a Web page in order to construct a two-way linked list, head node as the first title;
Step 2: Initialization *root.level* = 0, *root.cn* = 0;
Step 3: For each subtitle *subTitle$_i$* ∈*subTitleSet* (i = 1, 2, ..., n);
Step 4: If *subTitle$_i$* is the level of subheadings *root.cn*++,
root.child$_{cn}$ = *subTitle$_i$* , *subTitle$_i$.parent* = *root*
subTitle$_i$.level = *root.cn*;
Step 5: If *subTitle$_i$* is a second-level subheadings
While *subTitle$_i$.parent* and not the subtitle *subTitle$_i$.parent* not null
subTitle$_i$.parent = *subTitle$_i$.parent.parent*,
subTitle$_i$.parent.cn ++, *subTitle$_i$.parent.child$_{cn}$* = *subTitle$_i$*,
subTitle$_i$.level = *subTitle$_i$.parent.cn*,
Step 6: If *subTitle$_i$* is level 3 sub title
While *subTitle$_i$.parent* is not level 2 sub title and *subTitle$_i$.parent* not null
subTitle$_i$.parent = *subTitle$_i$.parent.parent*,
subTitle$_i$.parent.cn++, = *subTitle$_i$.parent.child$_{cn}$* = *subTitle$_i$*,
Step 7: Return *root*;
End of algorithm

Table 5: Distribution of web pages

Type of web page	Computer network	Operating system
Biadu baike web page	323	153
Genera; web]age	359	197

Table 6: Performance of document logical structure tree construction

	Baidu baike web page		General web page	
Method	Computer	Operatine	Computer	Operatine
M1	0.196	0.250	0.320	0.498
M2	0.169	0.241	0.252	0.452
M3	0.114	0.184	0.232	0.365
M4	0.066	0.128	0.226	0.300

EXPERIMENTATION

Database consists of web pages which downloaded from Baidu baike and general sutes. The dis tribution of web pages is shown in Table 5.

Fivefold cross vaildation on the database is conducted and average error rate adopted to evaluate the performance of different methods (Bille 2005). The performance of document logical structure tree consitruction is shown in Table 6, where M1, M2, M3 and M4 denote the method of "with the level of sub-title recognition title", "Logical constraint-based conflict detedtion and correction" "K-order optimal sequence solution" and " Based on solved the sub-structure of the optimal sequence", respectively.

CONCLUSION

This chapter around the Chapter Web page document logical structure extraction workflow expands the description, including the following aspects.

First, data preprocessing, including Web page noise filtering, candidate generation and data of the sub-title label work and pretreatment formal description and data preprocessing algorithm (Wei *et al*., 2010b).

Second, with the level of the sub-title identification phase of the mission, given the solution ideas, including the task of feature selection with the said program to develop model training and model predictions (Wei *et al*., 2012b; Wei and Qi, 2011).

Third, for the optimal sequence of sub-heading level to solve the problem, first give the formal description, followed by easy-to-digest proposed three solutions. Conflict detection and correction algorithm based on logical constraints from the point of view of the observation of the data set presented several logical constraints sub-heading level to detect the presence of conflicting and correct. K-order optimal sequence algorithm K-order characteristics to represent the context of the current sub-title, narrow understanding of the spatial extent of the threshold at the same time. Based on the sub-structure of the optimal sequence algorithm with a minimum semantic logic block to represent Web pages in separate units, with richer content than the sub-title while solving algorithm based on the consideration of the sub-structure of the tree in order K optimal sequence information (Wei and Zhou, 2012).

Fourth, a detailed description with a level of sub-headings in the document logical structure tree construction algorithm (Wei and Ma, 2012).

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their valuable comments. This program is supported by Scientific Research Program Funded by Shaanxi Provincial Education Department (Program 2010JK723) and Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2012JM8047). This program is and Supported by China Postdoctoral Science Foundation (No. 2013M542370). And this project is also supported by NSFC Grant (Program No. 60825202, 60803079, 11226173). It is also supported by National High-Tech Research and Development Plan of China under Grant No. 2008AA01Z131 and by Science and Technology Project of Xi'an (CX1262⑨) and Scientific Research Program Funded by Xi'an University of Science and Technology (Program No. 201139) and supported by the Specialized Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20136118120010).

REFERENCES

Amari, S. and S. Wu, 1999. Improving support vector machine classifiers by modifying kernel functions. Neural Networks, 12: 783-792.

Ashish, N. and C. Knoblock, 1997. Wrapper generation for semi-structured Internet sources. SIGMOD Rec., 26: 8-15.

Bille, P., 2005. A survey on tree edit distance and related problems. Theor. Comput. Sci., 337: 217-239.

Cauwenberghs, G. and T. Poggio, 2001. Incremental and Decremental Support Vector Machine Learning. In: Advances in Neural Information Processing Systems 13, Leen, T.K., T.G. Dietterich and V. Tresp (Eds.). MIT Press, Cambridge, UK., ISBN: 978026212241

Dumais, S., 1998. Using SVMs for text categorization. IEEE Intell. Syst., 13: 21-23.

EL-Shayeb, M.A., S.R. El-Beltagy and A.A. Rafea, 2009. Extracting the Latent Hierarchical Structure of Web Documents. In: Advanced Internet Based Systems and Applications, Damiani, E., K. Yetongnon, R. Chbeir and A. Dipanda (Eds.). Springer, USA., pp: 305-313.

Joachims, T., 2002. Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. Springer, USA.

Kurland, O. and L. Lee, 2005. PageRank without hyperlinks: Structural re-ranking using links induced by language models. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 2005, Salvador, Brazil, pp: 306-313.

Lewis, D.D., 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. Proceedings of the 10th European Conference on Machine Learning Chemnitz, Germany, April 21-23, 1998, Springer Berlin, Heidelberg, pp: 4-15.

Manevitz, L.M. and M. Yousef, 2001. One-Class SVMs for document classification. J. Mach. Learn. Res., 2: 139-154.

Mao, S., A. Rosenfeld and T. Kanungo, 2003. Document structure analysis algorithms: A literature survey. Proceedings of the SPIE Document Recognition and Retrieval X, January 20, 2003, USA., pp: 197-207.

McCallum, A. and N. Kamal, 1998. A comparison of event models for naive bayes text classification. Proceedings of the Workshop on Learning for Text Categorization, July 26-27, 1998, AAAI Press, Madison, Wisconsin, USA., pp: 41-48.

Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn., 1: 81-106.

Quinlan, J.R., 1996. Learning decision tree classifiers. ACM Comput. Surv., 28: 71-72.

Rish, I., 2001. An empirical study of the naive bayes classifier. Proceedings of the IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence, August 4, 2001, Seattle, USA., pp: 41-46.

Salton, G., A. Wong and C.S. Yang, 1975. A vector space model for automatic indexing. Commun. ACM, 18: 613-620.

Wei, W., A. Gao, B. Zhou and Y. Mei, 2010a. Scheduling adjustment of mac protocols on cross layer for sensornets. Inform. Technol. J., 9: 1196-1201.

Wei, W., B. Zhou, A. Gao and Y. Mei, 2010b. A new approximation to information fields in sensor nets. Inform. Technol. J., 9: 1415-1420.

Wei, W. and Y. Qi, 2011. Information potential fields navigation in wireless *Ad-Hoc* sensor networks. Sensors, 11: 4794-4807.

Wei, W., H. Yang, H. Wang, R.J. Li and W. Shi, 2011. Queuing schedule for location based on wireless ad-hoc networks with d-cover algorithm. Int. J. Digital Content Technol. Appl., 5: 356-363.

Wei, W. and L. Jun, 2012. The integration of p-Laplace model certifiable protocols with ID-based group key in WSNs. Int. J. Digital Content Technol. Appl., 6: 364-372.

Wei, W. and H. Ma, 2012. ARMA model and wavelet-based ARMA model application. Applied Mech. Mater., 121-126: 1799-1803.

Wei, W. and B. Zhou, 2012. A p-Laplace equation model for image denoising. Inform. Technol. J., 11: 632-636.

Wei, W., X.L. Yang, B. Zhou, J. Feng and P.Y. Shen, 2012a. Combined energy minimization for image reconstruction from few views. Math. Problems Eng., 10.1155/2012/154630

Wei, W., X.L. Yang, P.Y. Shen and B. Zhou, 2012b. Holes detection in anisotropic sensornets: Topological methods. Int. J. Distrib. Sensor Networks, 10.1155/2012/135054

Weston, J., S. Mukherjee, O. Chapelle, 2001. Feature Selection for SVMs. In: Advances in Neural Information Processing Systems 13, Leen, T.K., T.G. Dietterich and V. Tresp (Eds.). MIT Press, USA., pp: 668-674.

Witten, I.H. and E. Frank, 2005. Data Mining Practical Machine Learning Tools and Techniques. 2nd Edn., Morgan Kaufman, San Francisco, CA., USA.

Zhang, H.P., H.K. Yu, D.Y. Xiong and Q. Liu, 2003. HHMM-based Chinese lexical analyzer ICTCLAS. Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing, July 11-12, Sapporo, Japan, pp: 184-187.