Journal of Theoretical and Applied Information Technology
30th November 2012. Vol. 45 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A HVS MODEL FOR REPRESENTATION OF DOMAIN-ORIENTED WEB PAGE TOPIC FEATURES

XIANGHUA WU, QIAO GUO, LEI LA, QIMIN CAO

School of Automation, Beijing Institute of Technology, Beijing 100081, China

ABSTRACT

Domain-oriented web page extraction is a new and practical direction in the field of information extraction. The paper focuses on the representation of domain-oriented web page topic features, and hierarchical vector space (HVS) model is put forward. Considering the hierarchical characteristics of the web page itself, topic features of the web page are expressed more effectively by HVS model from the facets of the page structure and the content. Then the topic-related page identification problem is discussed by the similarity calculation. Experimental results show good accuracy and applicability for our system to domain-oriented web extraction.

Keywords: *Domain-Oriented, Hierarchical Vector Space Model, Information Extraction*

1. INTRODUCTION

With the rapid development of Internet, web has become the largest and most widely known repository of information in the world. Some types of web-based information extraction(IE) technology are emerged, which have a very wide range of applications, such as natural language processing[1, 2], knowledge discovery[3] and data mining[4]. Nowadays, according to the theoretical basis of IE technology, the methods of web information extraction are generally divided into four categories: IE based on natural language processing, statistical learning-based IE, IE based on ontology, IE based on html structure tree.

However, with the requirements of various information services gradually increased, the universal web extraction methods have been unable to meet the personalized needs. Web pages can be collected as much as possible but collection accuracy is relatively not high. Experimental results have shown that even for large-scale web extraction system, the coverage of web is only 30 percent to 40 percent. Moreover, conventional web system refreshes one time will take several weeks to a month[5,6]. In contrast, web information extraction methods for a specified domain can automatically search the pages related to the field and have no need to traverse the entire network, which significantly improved the effectiveness of information extraction and effectively shortened the update cycle.

This paper mainly researches the representation of domain-oriented web page features. Based on conventional vector space model, we propose a novel feature representation method called hierarchical vector space(HVS) model. Considering the hierarchical characteristics of the web page itself, the topic features of the web pages are expressed with a multi-dimensional vector from the facets of the page structure and the content. Then the topics of the domain can be established if we select the sample web pages in a specified field. Next we search the web pages related to the topic, which features have higher similarity with the topics. Based on the above analysis, the topic related web pages can be efficiently searched. Therefore, the representation of characteristics for the web pages in a specified field is a critical issue, which is also the focus of this paper.

2. FEATURE OF WEB PAGES

Based on general web extraction technology, domain-oriented web information extraction methods also take use of web page parsing technology and page text classification methods to satisfy effective web extraction. Overall, the web extraction system for a specified field consists of three parts: the domain topic training process, web searching process and page text classification process. System architecture diagram is shown in Figure1.

Journal of Theoretical and Applied Information Technology
30th November 2012. Vol. 45 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

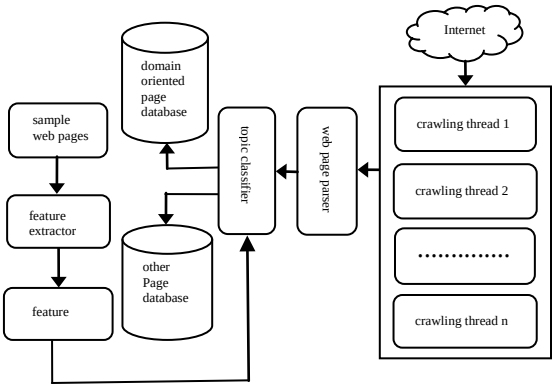

Figure 1: System Architecture Diagram

2.1 Web Extraction Principle

Domain-oriented web extraction method works normally around an explicit topic. Therefore, when we want to extract web pages in a field, firstly we should select some pages as sample pages from the authoritative websites related to the field. Researches have shown that the authoritative web sites related to one topic are easy to be found and URL of the pages on the same topic in one site tend to have a high level of similarity. So the design of the search strategy based on the following two assumptions: (1) manually selected initial websites and the search scope is limited in the same site once the sample pages are specified. (2) the web pages structure can be classified by the similarity of the URL.

2.2 Feature Of Web Page Stru-Cture

If we aim at extracting the topic-oriented web pages, it is critical to calculate the feature similarity of the test page and sample page. So how to define meaningful topic and how to express the practical theme of the system should be solved as the basic problem.

A web page contains different types of information areas. Normally, it is composed of five functional areas: the website header, text area, advertising area, navigation area and copyright area. Generally, text area is the part we are most concerned about.

Each website has unique design style, and location of each functional area in a page may not be exactly the same. Therefore, we cannot judge the area function only by its location. The vision tree is introduced by means of Vision-Based Page Segmentation (VIPS)[7]. Then we establish a vision tree modeling and the page might is divided into some blocks. Furthermore, we represent the feature of web page through the characteristics of the blocks and their weights.

By means of web Cascading Style Sheets (CSS) [8], part attributes of web page structure are shown in Table Ⅰ to Table Ⅲ. They are CSS font properties , text properties and background attribute. The feature attributes of the respective blocks can be obtained by training them, which reflect the characteristics of each functional block. e.g., background for website header is more complicated than the text area, and the font styles are relatively complex. In different blocks, weight of the same attribute is not the same. Accordingly, the weight of background in website header is higher than in the text area.

Table Ⅰ: Attribute Features Of Font

Attribute	Value
font-type	{"1pt","2pt","3pt",...}
font-size	{number|inherit|medium|large|larger|x-large|xx-large|small|smaller|x-small|xx-small}
font-weight	{ normal|bold|bolder| lighter | normal }
font-style	{ inherit| italic| normal| oblique}
font-variant	{normal| small-caps}

Table Ⅱ: Attribute Features Of Text

Attribute	Value
text-decoration	{inherit| none| underline| overline | line-through | blink}
text-transform	{capitalize| uppercase| lowercase| none}
text-align	{ left| right| center| justify }
white-space	{ normal| pre| nowrap }

Table Ⅲ : Attribute Features Of Background

Attribute	Value
background-color	number
background-image	{url (URL) |none}
background-repeat	{inherit| no-repeat| repeat| repeat-x | repeat-y }
background-attachment	{fixed| scroll}
background-position	{left| right| top| bottom| center| number}

2.3 Feature Of Web Page Text

The content feature of different functional blocks is not the same, with which we can identify the functional block. For instance, the number of pictures is not the same in different blocks, as well

Journal of Theoretical and Applied Information Technology
30th November 2012. Vol. 45 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

as the size of the picture. There are more pictures in the advertising area than in other areas. Moreover, the hyperlink number of the different blocks will also not the same. Number of hyperlink in navigation area is more than in the text area. Therefore feature representation of functional blocks are performed further by using these semantic features. In Table Ⅳ to Table Ⅷ, the specific features of contents are given.

Table Ⅳ: Content Features Of Images

Feature	Meaning
image-number	number
image-size	Img-size

Table Ⅴ: Content Features Of Hyperlink

Feature	Meaning
hyperlink-number	number
hypertext-size	Hypertext-size

Table Ⅵ: Content Features Of Text

Feature	Meaning
text-length	Inner Text-length
text-keywords	Inner Text keywords {"copyright"}

Table Ⅶ: Content Features Of Interaction

Feature	Meaning
interaction-number	interaction-Num
interaction-size	interaction-Size

Table Ⅷ: Content Features Of Form

Feature	Meaning
form-number	form-Num
form-size	form-Size

3. REPRESENTATION OF FEATURE VECTOR

We represent the feature vector by means of Vector Space Model. By means of VSM, document d

is mapped to vector $V(d)$,which is defined below:

$$V(d) = (t_1, \omega(t_1); t_2, \omega(t_2); ...; t_n, \omega(t_n)) \quad (1)$$

where $t_i (i = 1, 2, ..., n)$ is a list of different items. $\omega(t_i)$ is the weight of t_i in d, and generally defined the frequency function of t_i in d.

3.1 Hierarchical Vector Space Model

Taking into account that the web page is a semi-structured document, based on VIPS, the features of a web page can be expressed with a multi-dimensional vector from the facets of the page structure and the content. We propose a novel feature vector representation called hierarchical vector space (HVS) model, which improves the conventional vector space model in the level structure. Considering the web page features of hierarchical relationships, the feature of the page is expressed more accurately with HVS model.

We represent topic feature of web page P with a multi-dimensional vector F_p that is composed of structural vector F_p^s and content vector F_p^c. As follows:

$$F_p = (F_p^s, F_p^c);$$

$$F_p^s = (f_1^s, f_2^s, f_3^s, f_4^s, \cdots);$$

$$F_p^c = (f_1^c, f_2^c, f_3^c, f_4^c, f_5^c, f_6^c, ...);$$

$$f_i^s = (f_{i1}^s, \omega(f_{i1}^s); f_{i2}^s, \omega(f_{i2}^s); \cdots; f_{im}^s, \omega(f_{im}^s); \cdots), \quad (2)$$
$$(i = 1, 2, 3, 4, \cdots);$$

$$f_i^c = (f_{i1}^c, \omega(f_{i1}^c); f_{i2}^c, \omega(f_{i2}^c); \cdots; f_{im}^c, \omega(f_{im}^c); \cdots)$$
$$(i = 1, 2, 3, 4, \cdots);$$

In(2), $f_{im}^s (m = 1, 2, 3 \cdots)$ and $f_{im}^c (m = 1, 2, 3, \cdots)$ respectively reflect structure feature vector and content feature vector, which corresponds the CSS attribute in Table Ⅰ to Table Ⅲ and content feature in Table Ⅳ to Table Ⅷ. Correspondingly, $\omega(f_{im}^s)$ and $\omega(f_{im}^c)$ respectively represent the weights of f_{im}^s and f_{im}^c. Then we describe the feature vector by means of Term Frequency-Inverse Document Frequency(TF_IDF).

Take f_3^c for example, it can be expressed as follows:

$$T_c = \{t \mid \forall t \in d_i, \forall d_i \in D\} \quad (3)$$

In above equation, t represents the word of block d_i in document D. If a word belongs to one block or the most blocks, it is described to have little impact or little weight. In contrast, the word is the topic key word we need. tw_c describes weight of the key word according to the frequency of the word in a block and can be defined as below:

$$tw_c(d_c, t) = tf_c(d_c, t) \cdot \log \frac{n}{df_c(t)}, \quad (4)$$

In(4), $f_c(d_c, t)$ reflects the frequency of t in D.

Journal of Theoretical and Applied Information Technology
30th November 2012. Vol. 45 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

To deal with the web functional blocks with variant lengths, we standardize each vector. Then the mode of the vector equals 1. f_3^c can be described as below:

$$\left(\frac{tf_{31}^c \cdot \log \frac{n}{df_{31}^c}}{\sqrt{\sum_{i=1}^{m} (f_{3i}^c \cdot \log \frac{n}{df_{3i}^c})^2}}, \frac{tf_{32}^c \cdot \log \frac{n}{df_{32}^c}}{\sqrt{\sum_{i=1}^{m} (f_{3i}^c \cdot \log \frac{n}{df_{3i}^c})^2}}, \ldots, \frac{tf_{3m}^c \cdot \log \frac{n}{df_{3m}^c}}{\sqrt{\sum_{i=1}^{m} (f_{3i}^c \cdot \log \frac{n}{df_{3i}^c})^2}} \right) \quad (5)$$

If $\| f_3^c \| = 1$, f_p^c can be expressed as(6):

$$\left(\bigcup_{i=1}^{m_1} \frac{tf_{1i}^c \cdot \log \frac{n}{df_{1i}^c}}{\sqrt{\sum_{i=1}^{m_1} (tf_{1i}^c \cdot \log \frac{n}{df_{1i}^c})^2}}, \bigcup_{i=1}^{m_2} \frac{tf_{2i}^c \cdot \log \frac{n}{df_{2i}^c}}{\sqrt{\sum_{i=1}^{m_2} (tf_{2i}^c \cdot \log \frac{n}{df_{2i}^c})^2}}, \bigcup_{i=1}^{m_3} \frac{tf_{3i}^c \cdot \log \frac{n}{df_{3i}^c}}{\sqrt{\sum_{i=1}^{m_3} (tf_{3i}^c \cdot \log \frac{n}{df_{3i}^c})^2}}, \ldots \right) \quad (6)$$

3.2 Similarity Of Web Pages

We calculate the similarity of two pages after the topic features of pages have been represented in formula(2). We take two pages P_i and P_j for example, with the Cosine theorem, similarity of P_i and P_j can be described as following :

$$sim(P_i, P_j) = \frac{F_{p_i} * F_{p_j}}{\|F_{p_i}\| \cdot \|F_{p_j}\|} \quad , \quad (7)$$

Assuming P_i is a sample page, and we have expressed feature of P_i. If we will search the same topic page with the page P_i in the internet, the similarity of P_i and the test page called P_j should be computed. If $sim(P_i, P_j)$ reaches to a specified threshold θ, P_i is considered the topic-related page we are searching for.

3.3 Algorithm Description

We build domain-oriented web page database and other web page database. Both initialization are empty.

Step1. We calculate the feature value based on the need of domain-oriented information extraction.

(1) Manually specify the initial sample web sites and pages;

(2) Based on the Vision-Based Page Segmentation, the web page will be divided into some blocks, in which the structure feature and content features are different. We assign weight to every block.

(3) Calculate the feature by means of TF-IDF.

Step2. We collect the web pages by the web searching strategy.

Step3. The similarity of the test page and the sample page will be computed, then we can identify the web pages we need. If the similarity reaches to a threshold θ, the page can be stored in the domain-oriented web page databases.

The algorithm is described as follows in detail:

Input : web pages(include sample pages $P_i (i = 1, 2, 3, \cdots, m)$ and test pages $T_j (j = 1, 2, 3, \cdots, n)$)
Output: domain-oriented webpage database group

```
begin
  database group={};
  for ( i =1; i ≤ m ; i ++)
    represent feature of P_i : F_{P_i} = (F^s_{p_i}, F^c_{p_i})
    calculate the weight of F^s_{p_i} and F^c_{p_i}
  represent the topic feature of the domain F_p
  for ( j =1; j ≤ n ; j ++)
    represent feature of T_j : F_{T_i} = (F^s_{T_i}, F^c_{T_i})
    calculate sim(F_p, F_{T_i})
    if sim(F_p, F_{T_i}) > θ
      group=group ∪ { T_i };
end
```

4. EXPERIMENTAL ANALYSIS

In order to test validity of the method above, we take academic domain oriented web extraction for example. We select the journal paper web pages of part academicians in Chinese Academy of Sciences and Chinese Academy of Engineering, which were indexed since 2002 by SCI (Science Citation Index) database, EI (Engineering Index) database and IEL

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(IEEE/IEE, Electronic Library) database. Some will be taken as the sample pages, others as the test pages. According to the disciplines, the pages are divided into five domains: Maths, Geoscience, Biomedicine, Electronics and Computer science. First we train the sample page to obtain the features of each field, then we will search the web pages related to a field in test datasets. The test datasets and experimental results are shown in Table Ⅸ and Table Ⅹ.

Table Ⅸ: Datasets

		website		
		SCI	*EI*	*IEL*
domain	*Maths*	346	121	103
	Geoscience	198	143	45
	Biomedicine	204	187	79
	Electronics	174	478	340
	Computer science	156	313	184
Sample page Num		583	713	476
Test page Num		495	529	275
Total page num		1078	1242	751

Table Ⅹ: Experimental Results

		Experimental results		Test page num
		Recall (%)	*Precision (%)*	
Maths	*VSM*	83.8	80.7	284
	HVSM	88.7	87.2	
Geoscience	*VSM*	82.4	84.7	167
	HVSM	87.4	90.7	
Biomedicine	*VSM*	82.7	85.2	202
	HVSM	90.6	91.0	
Electronics	*VSM*	85.7	80.2	488
	HVSM	88.3	85.2	
Computer science	*VSM*	84.2	79.6	158
	HVSM	89.9	85.5	

From table Ⅹ, we can see that comparing with web extraction method by means of conventional VSM, our method with HVS model(HVSM in the table) achieves a better performance. Moreover, the recall ratio and precision of all category domains are more than 80 percent, even in several domains they are more than 90 percent. Experimental results show that with HVS model, more features of web structure and the content are explored, so extraction efficiency can improve remarkably. In addition, we found the precisions of domain about Electronics and Computer are relatively low, which is consistent with the actual situation. The two fields have relatively high similarity, therefore identification of web page becomes harder and precision correspondingly be lower than in other domains.

5. CONCLUSION

In this paper, we focus on the representation of web page topic features. Hierarchical vector space model is proposed. According to the hierarchical characteristics of the web page itself , the topic features of the web page are expressed accurately with a multi-dimensional vector from the page structure and the content. Then the topics of the domain can be expressed. Next we can search the suitable web pages in the domain. Experimental results about academic domain oriented area show that search method with HVS model has high efficiency and strong page adaptability. As a follow-up study, we should improve the maintenance work of the search method fault tolerance and object properties, making the search have more domain adaptability and the transplantation have more convenience and flexibility .

ACKNOWLEDGEMENTS

The paper is supported by China Association of Science and Technology.

REFERENCES:

[1] H. Cui, M. Kan, T. Chua, "Soft pattern matching models for definitional question answering", ACM Transactions on Information Systems, Vol. 25 , No. 2, pp. 1-30, 2007.

[2] E.Nyberg, T.Mitamura, J.Callan, et al, "The JAVELIN question-answering system at TREC 2003: a multi-strategy approach with dynamic planning", Proceedings of TREC 12, pp. 93-108, November, 2003.

[3] R J.Mooney, R.Bunescu, "Mining knowledge from text using information extraction", ACM SIGKDD Explorations Newsletter-Natural Language Processing and Text Mining, Vol. 7, No. 1, pp. 3-10, 2005.

[4] N.Kobayashi, R.Iida, K.Inui, et al, "Opinion mining on the web by extracting Subject -Attribute-Value Relations", Proceedings of the AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, pp. 470-481, 2006.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[5] CC.Aggarwal, F.AL-Garawi, Ps.Yu,"Intellig- ent crawling on the world wide web with arbitrary predicates", Proceedings of the 10th International Conference on World Wide Web, pp. 96 -105, 2001.

[6] S.Brin, L.Page, "The anatomy of a large-scale hypertextual web search engine", Proceedings of the Seventh International World Wide Web Conference, Computer Networks and ISDN Systems, Vol.30, No 1-7, pp.107-117, 1998.

[7] D.Cai, S.Yu, J.Wen, et al, "VIPS: a vision based page segmentation algorithm", Technical Report，MSR-TR-2003-79, Microsoft Research, 2003.

[8] G.Salton, A.Wong, C.S.Yang, "A vector space model for automatic indexing", Communications of the ACM, Vol. 18, No. 11, pp. 613-620, 1975.