A Web Site Classification Approach Based On Its Topological Structure[*]

Ji-bin Zhang，Zhi-ming Xu，Kun-li Xiu，Qi-shu Pan

School of Computer science and Technology，Harbin Institute of Technology
No.92,West Da-Zhi Street,Nangang District，Harbin 150006, China
Phn: +86-451-86414495, Fax: +86-451-86413309,
zjbxgg@hit.edu.cn,xuzm@hit.edu.cn

Abstract:

Automatic web site classification has a wide application prospect; however, there are few researches on it. Different from pure texts, web sites are the combination of a large number of web pages via hyperlinks, so text classification methods are not suitable to classify them directly. This paper proposes a web site classification approach based on its topological structure. Given a web site, firstly we represent its topological structure as a directed graph, and from which we extract a strongly connected sub-graph including the site's home page. Secondly, we use an improved PageRank algorithm on the sub-graph to select some topic-relevant resources, and represent them as a topic vector of the site. Finally we use an SVM classifier to classify the site in term of its topic vector. Some experiments are conducted for web site classification. Experimental results show our approach achieved better performance than traditional super page-based web site classification approach.

Key words:

web site classification,topological structure of web site,hyperlink analysis,topic vector of web site

[*] We thank all students of Natural Language Processing Research Center in Harbin Institute of Technology for discussing some issues about this paper. Research for this article was supported by the National Natural Science Foundation of China(60773070, 60736044).

1 Introduction

With the rapid development of Internet, the information of network grows explosively. According to the statistical data released by Google, Google currently has indexed over one trillion web pages and this figure is still rapidly increasing every day. Internet has already become the most important source of information and knowledge in scientific research, education and other fields. Due to its mass, variable, and non-semantic characteristics, it is not easy for people to find the information they want quickly and accurately. How to find the information we need from such a huge source has become an important objective we need to study.

For the moment, there are two kinds of services that can help us to retrieve information in the Internet: search engines like Google and directory services like Yahoo! and DMOZ. Search engines usually return some web pages matched with queries. However people sometimes need to find some web sites related with a certain subject. For example, when people want to buy something, they will try to find the retailer’s web sites instead of web pages which only contain descriptions of commodities. Directory services supply a navigation mechanic of web sites by collecting a number of web sites and manually classifying them into different directories. But they spend lots of manual editorial work to maintain directory services.

The technology of web information navigation, especially automatic classification of web information is becoming the research focus. Considering that automatic web site classification is significant to maintenance directory services, this paper mainly studies automatic web site classification approaches. Because a web site is the combination of a large number of web pages via hyperlinks, which has richer structure information than single web page, text classification approaches are not suitable to classify it directly. This paper proposes a web site classification approach based on its topological structure. Given a web site, firstly we represent its topological structure as a directed graph, and from which we extract a strongly connected sub-graph including the site’s home page; secondly, we use an improved PageRank algorithm on the sub-graph to select some topic-relevant resource, and represent them as a topic vector of this site; finally we use a SVM classifier to classify the site in term of its topic vector.

The rest of this paper is organized as follows. Section 2 gives a summarization of previous research on web page and web site classification. In section 3 we describe our web site classification approach. In section 4, we contact some experiments to test our web site classification approach, and the conclusion is given in the last section.

2 Related work

Automatic classification of web pages has been studied for a long time, some text classification algorithms like Naïve Bayes(McCallum 1998;Mitchell 1996), KNN(Lam 1998;Masand 1992), and SVM(Joachims 1998;Kwok 1998) have been successfully applied. Apart from the content of web pages, Chakrabart(Chakrabarti 1998) and Craven (Craven 1999) improved the accuracy of web page classification by introducing hyperlink analysis. However, there is a little research on web site classification, the difficulty in which is that a web site consists of many pages, and each page has its own topic, a site's topic can not be reflected by a single web page. A famous web site classification method is super page-based method (Ester 2002), which represents a web site as a single virtual web page combined by all its pages, and Pierre (Pierre 2001) improved it by introducing web pages' meta data, such as title, keyword, and so on. Terveen(Terveen 1999) represented a web site as a directed graph and combined content and hyperlink analysis to classify it. Ester(Ester 2002; Ester 2004) gave an empirical study on web site classification, and proposed several solutions of web site classification; on the basis of the research of Ester, Kriegel(Kriegel 2004) introduced a method that represented a web site as a topic-frequency vector. In addition, YongHong Tian(Tian 2004) used a multi-scale tree model to represent a web site, De-yu Fu proposed a key resource-based web site classification method(Fu 2006), and Bao-li Dong employed a hybrid vector space model to recognize the subject of web sites(Dong 2005).

3 Web site classification approach based on its topological structure

In this section, we mainly discuss our web site classification approach based on its topological structure. This approach mainly includes several phases: represent a web site's topological structure as a directed-graph, extract some topic-relevant resources from the site's topological graph, represent extracted topic-relevant resource as a topic vector, and use the topic vector to classify the site.

3.1 Representing a web site's topological structure as a directed-graph

In this section, we represent the topological structure of a web site as a directed graph. Some definitions about the directed graph are given as follows:

Definition 1. Directed graph: A directed graph is an ordered pair $D=<V, E>$. D represents the topological structure of a web site. V is a set of vertices, each vertex of V is a page; E is a set of directed edges, which is a subset of $V\times V$, each directed edge $e=(u, v)$ means a hyperlink e from page u to page v.

Definition 2. Degree: A directed graph $D=<V, E>$. For each vertex $v_i \in V$, the number of edges linked from V_i is defined as the out-degree of v_i, and the number of edges linked to v_i is defined as the in-degree of v_i. The sum of in-degree and out-degree of v_i is defined as the degree of $v_{i.}$

Definition 3. Path: A path in a directed graph $D=<V, E>$ from v_m to v_n is defined as a sequence of vertices $\{v_m, v_{m1}, v_{m2} \ldots , v_n \}$, which includes edges (v_m, v_{m1}), $(v_{m1}, v_{m2}) \ldots (v_{mi}, v_n)$.

Definition 4. Sub-graph: There are two directed graphs: $D=<V, E>$ and $D'=<V', E'>$, D' is called as a sub-graph of D if $V` \subset V$ and $E` \subset E$.

Definition 5. Strongly connected graph: A directed graph $D=<V, E>$ is called as a strongly connected graph if there is a path between any pair of two vertices.

Fig.1 is an example of a web site, where the site is represented as a directed graph, if page *A* can reach page *B* via inner hyperlinks, then there is a path from *A* to *B*. If any pair of pages in this graph has a path to connect them, then we call it strongly connected.

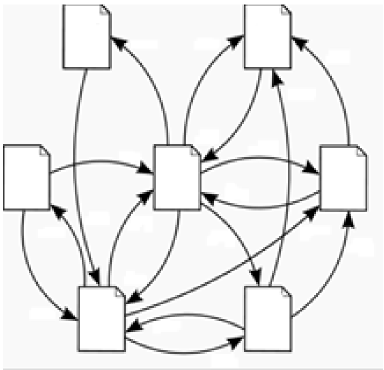

Fig.1 A topological graph of a web

3.2 Extracting topic-related resource from web site's topological graph

After the topological graph of a web site is built, we put the emphasis on extracting topic-related resource from it. According to some literatures' point of view, a web site's home page may be the most topic-relevant to the site(Dong 2005), and a site's pages with the same topic usually have a compact link structure(Liu 2006). In addition, web site designers generally hope that outgoing-linked pages should be topic-relevant to the current page, so we can assume a pair of pages in a site is topic-relevant if there is a hyperlink between them(Ester 2004). According to this assumption, we can infer that a pair of pages in a site should be topic-relevant if there is a path between them.

Considering the above all, we think that a site's topic-relevant resource should be located on a strongly connected sub-graph including the site's home page, on which we can use hyperlink analysis technology to select important topic-relevant sources.

The PageRank(Page 1999) algorithm is often used to compute the importance of web pages, which regards the entire web as a directed graph, and ranks pages through hyperlink analysis. This paper will use an improved PageRank algorithm in a site's sub-graph to rank pages to select important topic-relevant sources from it.

3.2.1 Improved PageRank Algorithm

PageRank is the earliest and the most successful algorithm applied to the hyperlink analysis on commercial search engines, which interprets a hyperlink from page A to page B as a vote, by page A, for page B. If pages that cast votes are important, they will make pages voted to be important. A simplified version of PageRank defined by Larry is as follows:

$$PR(s)=\sum_{i=1}^{N}\frac{PR(P_i)}{C(P_i)} \tag{1}$$

where *s* is a page, *PR*(*x*) means the rank score of page *x*, *N* is the in-degree of s, P_i is the page linked to *s*, and $C(P_i)$ is the out-degree of page P_i.

In formula (1), the rank score of page P_i is divided by its out-degree, and each page linked from P_i is distributed with the same rank score. There is a small problem with formula (1). Assumed that there are two or more pages linked to each other but to no other pages, and there is a hyperlink linked to one of them; after some iterations, rank scores are accumulated into them but never distributed out from them. This scenario is called rank sinking. To solve rank sinking problem, Larry modified the original PageRank formula as follows:

$$PR(s)=(1-d)+d\times\sum_{i=1}^{N}\frac{PR(P_i)}{C(P_i)} \tag{2}$$

where *d* is usually set as 0.85, it is the probability that users continue to view pages linked from the current page *s*, (1- *d*) is the probability that users leave the current page *s* and skip to other web pages.

In PageRank algorithm, each page *s* distributes its rank score to pages linked from *s* averagely. But the average distribution scheme of rank scores among pages is not suitable for the demand of web site classification. For a web site, we aim to select topic-relevant resource from its sub-graph, so we consider that the rank scores should be distributed according to page similarity. If one page A is more similar to pages linked to A, A will get more rank scores, otherwise it will get less. Here, we use an improved PageRank formula to distribute rank scores among pages, which is shown as follows(Yuan 2007) :

$$PR(s) = (1-d) + d \times \sum_{i=1}^{N} PR(P_i) \times \frac{sim(P_i, s)}{\sum_{j=1}^{Mi} sim(P_i, Q_{ij})} \qquad (3)$$

where $sim(P_i,s)$ is the page similarity, Q_{ij} is a page linked from P_i, M_i is the number of pages linked from P_i.

Fig.2 is an example of the improved PageRank formula. Assumed that page *A* has two hyperlinks linked to page *B* and page *C* respectively, *PR*(*A*) is 1, *sim*(*A*, *B*) is 0.8, and *sim*(*A*, *C*) is 0.4. The rank score distributed from *A* to *B* is 1*0.8/ (0.8+0.4) =0.6666, and *C* gets 1*0.4/ (0.4+0.8) =0.3333.

3.2.2 Computation of link-based page similarity

The improved PageRank formula uses the page similarity to distribute the rank scores among pages. In general, the similarity between pages can be computed according to their contents(Wang 2003). Considering the computation cost of content-based page similarity, we use the computation methods of link-based page similarity, which only analyze hyperlinks among pages instead of their contents.

In Fig.2, Page *A* has two hyperlinks linked to page *B* and page *C* respectively. According to Literature(Ester 2004), if *B* and *C* are both linked to or from the same page, they may have the same topic. The more are the pages linked to or from both *B* and *C*, they are more topic-relevant. In other words, they are more similar.

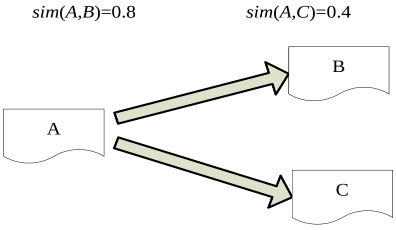

Fig.2 The improved PageRank formula

According to Literature (Wang 2003), we number all the pages in a site as {1, 2, 3 … n}; for each page *s*, we construct two vectors: V_s^{out}, V_s^{in}. If the i^{th} page has a hyperlink linked to *s*, then the i^{th} item of V_s^{in} is 1, otherwise it is 0. Similarly, if i^{th} page has a hyperlink linked from *s*, the i^{th} item of V_s^{out} is 1, otherwise it is 0. Considering the above all, Literature (Wang 2003) gave the in-link-similarity, out-link-similarity and link-based similarity between page *A* and page *B* as follows:

$$Similarity^{in}(A,B) = \frac{v_A^{in} \cdot v_B^{in}}{\left\| v_A^{in} \right\| \cdot \left\| v_B^{in} \right\|} \tag{4}$$

$$Similarity^{out}(A,B) = \frac{v_A^{out} \cdot v_B^{out}}{\left\| v_A^{out} \right\| \cdot \left\| v_B^{out} \right\|} \tag{5}$$

$$Similarity(A,B) = Similarity^{in}(A,B) \times Similarity^{out}(A,B) \tag{6}$$

In the above similarity formulas, the more common pages are linked to *A* and *B*, the bigger is *Similarity*^{in}(*A*,*B*); the more common pages are linked from *A* and *B*, the bigger is *Similarity*^{out}(*A*,*B*).

For a given web site, we firstly extract the strongly connected sub-graph including the site's home page, and then we use formulas (3) and (6) to compute the rank score of each page, rank these pages according to their rank scores, and finally select some high-scored pages as topic-relevant resource of the site.

3.3 Represent the Topic Vector

After ranking the pages in the sub-graph, some topic-relevant resources on the sub-graph are selected. Now we should consider how to represent extracted pages and their hyperlinks. According to literatures' point of view(Hodgson 2001), 61% anchor texts of hyperlinks can reflect the topic of pages they link to. So we view anchor text of hyperlinks as a site's structure feature of sites, and view content text of pages as content feature of sites. Under vector space model, we combine content feature and structures feature of a site to a mixed vector, called a topic vector, which is shown as follows:

$$v = (w_1^{'}, w_2^{'} \cdots w_m^{'}, w_1, w_2 \cdots w_n), \quad l = m + n \tag{7}$$

where v is a *l*-dimension mixed vector, $w_i^{'}$ is the weight of the structure feature term $t_i^{'}$ and w_j is the weight of the content feature term t_j. Here, we use Information Gain (IG) method to select content and structure feature items, and use traditional entropy weighting method to weight structure feature items (anchor text items).

$$a_{ik} = \log(TF_{ik} + 1) \times \left\{ 1 + \frac{1}{\log(N)} \sum_{j=1}^{N} \left[\frac{TF_{ik}}{n_i} \log\left(\frac{TF_{ik}}{n_i} \right) \right] \right\} \tag{8}$$

where a_{ik} is the weight of term i in the site k . TF_{ik} is the frequency of term i appearing in the site k. N is the number of all training sites. n_i is the numbers of sites which include term i. But when we want to weight content feature items, we should consider not only frequency information of terms but also their location information on pages. Some of HTML tags are important for reflecting topics of pages, such as "title", "keyword", and "description", and they generally summarize the content of pages. In addition, the titles, bold, italic information in the body of pages are also important to reflect the topic of pages. So we put our emphasis on considering the impact on pages' topics of a tag set, S= {title, keywords, descriptions, H1, H2, H3, B, U, I}, and enlarge the weights of terms which appear in tags of S. here we give an improved entropy weighting formula to weight them, which is shown as follows:

$$a_{ik} = \log\left(\sum_{\beta\in S} w^{\beta} \times TF_{ik}{}^{\beta} + 1\right) \times \Bigg\{ 1 + \frac{1}{\log(N)} \sum_{j=1}^{N} \Bigg[\frac{\sum_{\beta\in S} w^{\beta} \times TF_{ik}{}^{\beta}}{n_i} \log\Bigg(\frac{\sum_{\beta\in S} w^{\beta} \times TF_{ik}{}^{\beta}}{n_i} \Bigg) \Bigg] \Bigg] \Bigg\} \qquad (9)$$

where $TF_{ik}{}^{\beta}$ is the frequency of item i that appears in the site k and locates on the tag β. w^{β} is the weighted coefficient for the tag β, and let $W^{title} > W^{keyword} > W^{description} > W^{H1} > W^{H2} > W^{H3} > W^{U} > W^{I}$.

4 Experiments

In our web site classification experiments, we use Google's navigation site (http://daohang.google.cn/) as our data source, from which we download 1127 web sites data from 16 categories, use 760 web sites data as our training samples, and use 367 web sites data as our testing samples. We use SVM model as our web site classifier, and Information Gain method is used for feature selection; in addition, we use traditional entropy weighting method and the improved entropy weighting method to weight structure terms and content terms respectively. All the experiments were implemented in C++ and tested on a PC equipped with AMD Athlon 3600+ processor and 1 GB main memory.

In our web site classification experiments, we use super page-based web site classification method as the baseline system, in which we limit the numbers of each site's web pages under a maximum of 50; for our web site classification method based on its topological structure, we only select top 20 pages as each site's topic-relevant data.

Fig.3, Fig.4 and Fig.5 show the comparison of our web site classification method based on its topological structure with super page-based web site classification method on precision, recall and F1 value. Table 1 shows the comparison of these two web site classification methods on macro-averaging and micro-averaging values. Experimental results show that our method achieves much better performance than super page-based web site classification method. Macro-averaging and micro-averaging values can be increased nearly by 20% with our method compared with those with super page-based method.

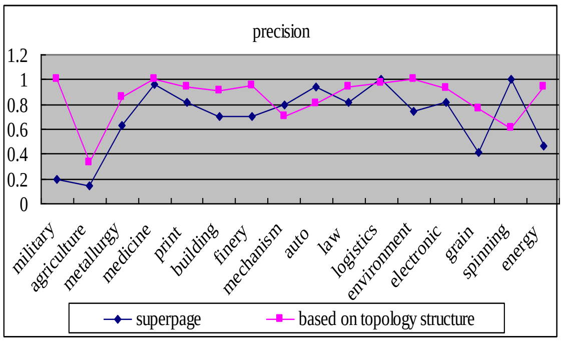

Fig.3 Comparison on precision

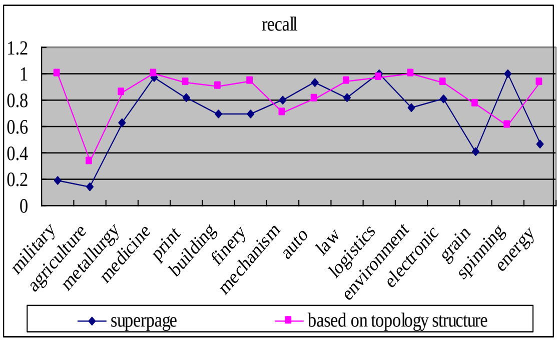

Fig.4 Comparison on recall

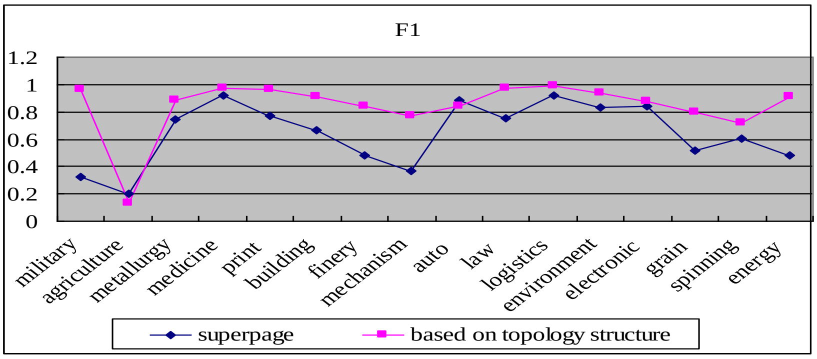

Fig.5 Comparison on F1

	based on topological structure	superpage
MacroP	0.847286	0.696746
MacroR	0.853727	0.669709
MacroF1	0.842088	0.643761
MicroP	0.882834	0.686649

Table 1 Comparison on macro-averaging and micro-averaging

To investigate the performance of the improved PageRank algorithm and traditional PageRank algorithm on web site classification, we conducted a comparison experiment for them. The experimental results are showed in Table 2. Although the improved PageRank algorithm decreases the MacroP than traditional PageRank algorithm, but it increases MacroR, MacroF1, and MicroP values evidently.

	Improved PageRank	PageRank
MacroP	0.847286	0.863231
MacroR	0.853727	0.80901
MacroF1	0.842088	0.814857
MicroP	0.882834	0.847411

Table 2 The effect of Improved PageRank

5 Conclusions

In this paper, we propose a web site classification approach based on its topological structure. The topological structure of a web site can be represented as a directed graph. Assumed that the topic-relevant resource of a web site is located on the strongly connected sub-graph including the site’s home pages, we use an improved PageRank algorithm based on link-based page similarity, which can efficiently rank pages in the sub-graph. For efficiently representing content feature and structure feature of a site, we mix them into a topic vector, and use an improved entropy weighting method to weight content terms according to their frequency and location information on pages. The experimental results of web site classification show that our web site approach can achieve better performance than traditional web site classification approaches.

6 References

Google: Search Engine. http://www.google.com/

Yahoo: rectory Service. http://www.yahoo.com/

DMOZ: Open Directory Project. http://DMOZ.org/

McCallum,A.and Nigam,K.,1998,A Comparison of Event Models for Naïve Bayes Text Classification, *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*.

Mitchell, T. M.,1996,*Machine Learning*. New York :McGraw Hill.

Lam,W. and Ho,C.Y.,1998,Using a Generalized Instance Set for Automatic Text Categorization, *proceeding of the 21st Ann International ACM SIGIR Conference on Research and Development n Information Retrieval Melboume, AU*, pp.81-89.

Masand,B., Lino,G. and Waltz,D.,1992,Classifying News Stories Using Memory Based Reasoning, *proceeding of the 15th Annual ACM SIGIR Conference*, Denmark: Copenhagen, pp. 59-65.

Joachims,T.,1998,The Categorization with Support Vector Machines: Learning with Many Relevant Features, *In European Conference on Machine Learning (ECML),*Chemnitz ,Germany, pp. 137-142.

Kwok,J.T.Y.,1998,Automatic Text Categorization Using Support Vector Machine, *Proceeding of International Conference on Neural Information Processing,* pp. 347-351.

Chakrabarti, S., Dom, B. and Indyk, P.,1998,Enhanced Hypertext Categorization Using Hpyerlinks, *Proceeding of the ACM SIGMOD Conference on Management of Data Seattle,* Washington, pp. 307-318.

Craven, M., DiPasquo, D., and Freitag, D., 1999,Learning to Construct Knowledge Bases from the World Wide Web, *In Artificial Intelligence*.

Ester, M., Kriegle, H.P., Schubert, M.,2002,Web Site Mining: A new way to spot Competitors, Customers and Suppliers in the World Wide Web, *Proceeding of 8th International Conference on Knowledge Discovery and Data Mining*.

Pierre, J. M.,2001,On the Automated Classification of Web Sites, *Linkoping Electronic Articles in Computer and Information Science* ,Vol. 6.

Terveen,L., Hill,W., and Amento, B.,1999, Constructing, Organizing, and Visualizing Collections of Topically Related Web Resources. *ACM Trans. on Computer-Human Interaction*,vol. 6,no.1,pp.67-94.

Ester, M., Kriegel, H.P., Schubert,M.,2004,Accurate and Efficient Crawling for Relevant Websites, *Proceedings of the Thirtieth international conference on Very large data bases*, Aug, pp.396-407.

Kriegel, H.P., Schubert, M.,2004,Classification of Websites as Sets of Feature Vectors, *proceedings of the IASTED International Conference DATABASES AND APPLICATIONS* ,Feb 17-19.

Tian, Y.H., Huang, T.J.,and Gao, W.,2004, A Web Site Representation and Mining Algorithm using a Multiscale Tree Model. *Journey of Software*, vol.15,no.9,pp.1393-1404.

Fu, D.Y., Dai, C.Q., and Zhong, W.,2006, A Web Site Categorization System Based on Key Resources, *Journey of Harbin Institute of Technology*, vol.38,no.1, pp.19-22.

Dong, B.L.， Qi, G.N，and Gu, X.J.,2005, Specific website subject recognition based on the hybrid vector space model. *Journal of Tsinghua university (Sci & Tech)* , vol.45,pp.1795-1801.

Liu, Y., Wang, B., Yang, Z.F., and Zhang, X., 2006,Link Analysis in Web Key Resources Discovery, *Proceedings of CNCCL*,pp.945-500.

Page, L., Brin, S., and Motwani, R., 1999,The PageRank Citation Ranking: Bringing order to the Web, Technical report, *Stanford Digital Libraries SIDL-WP-1999-0120*.

Hodgson, J., 2001,Do HTML Tags Flag Semantic Content? *IEEE Internet Computing*, vol. 5,no.1,pp.20-25.

Wang, X.Y.，Xiong, F.，Ling, B.，and Zhou, A.Y.,2003,A Similarity-Based Algorithm for Topic Exploration and Distillation, *Journey of Software*,vol.14,no.09, pp.1578-1583.

Yuan, F.Y.,and Zhang, Y.Y.,2007,The research and improvement of relevance ranking method based on link analysis, *Computer Engineering and Design*,vol.28, no.7,pp.1630-1631.