WEB PAGE CLASSIFICATION CONSIDERING PAGE GROUP STRUCTURE FOR BUILDING A HIGH-QUALITY HOMEPAGE COLLECTION

Yuxin Wang and Keizo Oyama
National Institute of Informatics, Hitotsubashi 2-1-2, Chiyoda-ku, Tokyo 101-8430, Japan

Keywords: Surrounding page group, three-way classification, recall and precision, quality assurance.

Abstract: We propose a web page classification method for creating a high quality homepage collection considering page group structure. We use support vector machine (SVM) with textual features obtained from each page and its surrounding pages. The surrounding pages are grouped according to connection type (in-link, out-link, and directory entry) and relative URL hierarchy (same, upper, or lower); then an independent feature subset is generated from each group. Feature subsets are further concatenated to compose the feature set of a classifier. The experiment results using ResJ-01 data set manually created by the authors and WebKB data set show the effectiveness of the proposed features compared with a baseline and some prior works. By tuning the classifiers, we then build a three-way classifier using a recall-assured and a precision-assured classifier in combination to accurately select the pages that need manual assessment to assure the required quality. It is also shown to be effective for reducing the amount of manual assessment.

1 INTRODUCTION

The web is becoming more and more important as a potential information source for adding value to high-quality scholarly information services. What is required here is a web page collection with guaranteed quality (i.e., recall and precision); however, the task of building such a collection is thought to demand a large amount of human effort because of the diversity in style, granularity and structure of web pages, the vastness of the web data, and the sparseness of relevant pages.

Many researchers have investigated classification of web pages, etc.; however, most of their approaches are of the best-effort type and pay no attention to quality assurance. Thus, we are trying to devise a method to build a homepage collection efficiently with high quality level assured.

We approach this problem in two steps: improving web page classification performance by exploiting information in surrounding pages considering page group structure, and accurately selecting web pages requiring manual assessment.

For the first step, we propose a method using support vector machine (SVM) with features that are obtained from the textual contents in each page and its surrounding pages considering their page group structure. For the second step, we build a three-way classifier using a recall-assured classifier and a precision-assured classifier in combination. We can exploit the first approach's potential performance by tuning each classifier independently.

We evaluated the proposed method with experiments using 'ResJ-01', a Japanese data set and 'WebKB', an English data set.

We intend to use the resulting web page collection for extending a guarantee-type information service such as CiNii (CiNii), not only as link targets but also as reference data for record linkage (Aizawa and Oyama, 2005) among papers, citations, and/or project reports. Considering the applications, 95% recall and 99% precision, for example, should be assured.

The rest of the paper is organized as follows. Related work is introduced in Section 2. The scheme of the three-way classifier is presented in Sections 3 and the proposed feature is explained in Section 4. Section 5 presents the experiment results and section 6 discusses the effectiveness of the proposed method. Finally, we conclude our work in Section 7.

Wang Y. and Oyama K. (2007).
WEB PAGE CLASSIFICATION CONSIDERING PAGE GROUP STRUCTURE FOR BUILDING A HIGH-QUALITY HOMEPAGE COLLECTION.
In *Proceedings of the Third International Conference on Web Information Systems and Technologies - Web Interfaces and Applications*, pages 170-175
DOI: 10.5220/0001271701700175

2 RELATED WORK

The method proposed in this paper belongs to the web page classification domain, and is closely related to the web page search and clustering domains. In these domains, what information sources to use is the first factor to be considered and how to use them is the second.

The previous studies tried to exploit, besides textual content, various web-related information sources such as html tags (Sun et al., 2002), URLs (Kan and Thi, 2005), sub-graphs of web pages (Sun and Lim, 2003; Masada et al., 2005), directory structure (Sun and Lim, 2003), anchor texts (Sun et al., 2002), contents of globally link-related pages (Glover et al., 2002), and contents of local surrounding pages (Wang and Oyama, 2006; Masada et al., 2005; Sun and Lim, 2003). All of these information sources except the last one are used to capture the features that are characteristic to the target pages, and are effective for selecting highly probable pages. The last one, in contrast, is used to collect information scattered over some linked pages so that potential target pages can be comprehensively gathered; however, it also tends to increase noise. Since comprehensiveness is a key factor for quality assurance of a web page collection, we mainly focus on the last one, i.e., the contents of local surrounding pages as information sources, in the current work.

Sun et al. (Sun et al., 2003) studied on exploiting the surrounding pages. They first classify each page only based on its content and then combine the results from other information sources, such as the link structure and directory structure. However, this approach will not work when an entry page contains no textual information except hyperlinks. Other studies first cluster web pages based on the local link structure and so on, and then merge the score (or weight) of each word to generate a document vector (Wang and Oyama, 2006; Masada et al., 2005). However, the effectiveness of this approach is limited, probably because it also merges many irrelevant words from the surrounding pages.

We also exploit the contents in surrounding pages considering page group and local link structure, but with a different approach. To reduce noise, we divide the surrounding pages to groups according to the structural relation with the target page, extract features from each group, and concatenate (rather than merge) the features from the groups, so that the contexts corresponding to the structural relation can be represented. We also propose a very simple method for exploiting html tag related information.

In addition, almost no previous studies tried to assure the quality level that would be required for practical applications. We are trying to solve this problem by building a three-way classifier using a recall-assured classifier and a precision-assured classifier in combination.

3 CLASSIFICATION SCHEME

Figure 1 shows the scheme of our classification method (95% recall and 99% precision are used as the quality requirement for illustration). We use two component classifiers to construct a three-way classifier. The recall-assured (precision-assured) classifier assures the target recall (precision) with the highest possible precision (recall).

All web pages are first input to the recall-assured classifier and its negative predictions are classified as "*assured negative*". The rest are then input to the

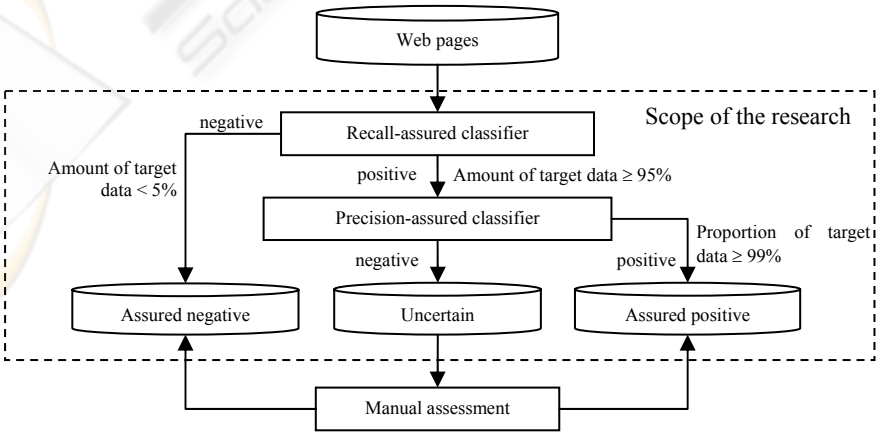

Figure 1: Scheme of the Classification.

precision-assured classifier and its positive predictions are classified as "*assured positive*". The remaining pages are classified as "*uncertain*", and these require manual assessment in order to keep the required quality. The target precision is assured in the "*assured positive*" output, the target recall is assured in the "*assured positive*" and "*uncertain*" outputs together, and the "*uncertain*" output should be minimized.

4 CLASSIFIER COMPOSITIONS

The support vector machine (SVM) is effective for text classification. We used the SVM^{light} package developed by Joachims with the linear kernel and tuned it with options *c* and *j*.

4.1 Surrounding Page Group and Feature Set

In the current work, given a page to be classified (current page), a surrounding page of the current page is defined as a page related to the current page by either of three connection types *c*: *in* (in-link), *out* (out-link), and *ent* (directory entry) in the directory path or in the directory subtree of the current page. We assign surrounding page groups $G_{c,l}$ to each surrounding page according to the connection type *c* and URL hierarchy level *l* (*same*, *upper* and *lower*) relative to the current page. The current page alone constitutes an independent group G_{cur}. All defined surrounding page groups are shown in Table 1. In view of a logical page group, each group has its own implicit meaning. For example, $G_{in,lower}$ consists of in-link pages in lower directories which might represent component pages having back links to the entry page, and $G_{ent,upper}$ consists of directory entry pages in upper directories which might represent the organization the entity of the current page belongs to.

We use textual feature $f_{t,v}(g,w_t)$, where *t* indicates a text type *plain* (plain-text-based) or *tagged* (tagged-text-based), *v* indicates a value type *binary* or *real*, *g* denotes a surrounding page group, and $w_t \in W_t$ denotes a feature word. A feature set is composed by concatenating one or more feature subsets $F_{t,v}(g) = \{ f_{t,v}(g,w_t) \mid w_t \in W_t \}$ for each *g* in G_{cur} and any number of $G_{c,l}$'s. For instance, the feature set "u-1" shown in Section 5 consists of feature subsets on G_{cur} and $G_{*,upper}$ (surrounding page groups of upper hierarchy level), the feature set "o-i-e-1" consists of feature subsets on G_{cur} and $G_{*,*}$ (all surrounding page groups), and the feature set "all_merged" consists of feature subsets on G_{cur} and $G_{all,all}$ (all the surrounding pages merged together).

Table 1: Surrounding page groups.

Hierarchy level *l*	Connection type *c*				
	r	*out*	*in*	*ent*	merged
same	G_{cur}	$G_{out,same}$	$G_{in,same}$	$G_{ent,same}$	$G_{all,same}$
lower		$G_{out,low}$	$G_{in,low}$		$G_{all,low}$
upper		$G_{out,upper}$	$G_{in,upper}$	$G_{ent,upper}$	$G_{all,upper}$
merged		$G_{out,all}$	$G_{in,all}$	$G_{ent,all}$	$G_{all,all}$

4.2 Text Type and Feature Word

We use two kinds of textual features and use *Chasen* for Japanese data set and *Rainbow* for English data set to tokenize the feature words from the texts.

Plain-text-based features $f_{plain,*}(*,*)$ are extracted from textual content excluding tags, scripts, comments, etc. We use top 2,000 feature words W_{plain} ranked by mutual information (Cover and Thomas, 1991).

Tagged-text-based features $f_{tagged,*}(*,*)$ are extracted from "*text*" segments that match either ">*text*<" or "<img…alt= "*text*"...>" and that are not more than 16 bytes long omitting spaces for the Japanese data set and not more than 4 words long for the English data set. The obtained words with not less than 1% file frequency for Japanese data set and all the obtained words for English data set are used as feature words W_{tagged}. We use the tagged-text-based features because the *text* segmets are considered to frequently contain property words.

In the experiments, a feature set is composed by using feature words either *plain* alone or *plain* and *tagged* together. The latter case is indicated by the suffix "_tag" of the run name.

A *binary* value $f_{*,binary}(g,w_t)$ represents the presence of w_t in *g*. A *real* value $f_{*,real}(g,w_t)$ represents the proportion of pages containing w_t within *g*. The *real* value is tested to see if the feature word distribution within surrounding page groups is informative. The two value types are exclusively used for composing a feature set. Use of the value type *real* is indicated by the suffix "_real" in the run name.

5 EXPERIMENTS

5.1 Experiments on ResJ-01 Data Set

We manually prepared 'ResJ-01', a Japanese data set of researchers' homepages consisting of 480 positive samples and 20,366 negative samples. *Five-*

fold cross validation was used. A feature set only composed of $F_{plain,binary}(G_{cur})$ was used as the baseline. Fig. 2 shows the overall performances and Fig. 3 and 4 zoom up high precision and high recall areas of the two best performing classifiers. Their performances compared with the baseline and all_merged are listed in Table 2. The results show that, in the high precision area, both o-i-e-1_tag_real and u-1_tag outperform the others; while in the high recall region, o-i-e-1_tag_real evidently outperforms all the others and u-1_tag also performs rather well.

5.2 Experiments on WebKB Data Set

We used the WebKB data set for testing the effectiveness of the proposed features. It is an English data set containing 8,282 pages collected from the computer science departments of 4 major universities and of other universities, and manually classified into 7 categories. 4 categories, i.e., *student*, *faculty*, *course*, and *project*, were used for our experiments. The pages of the respective categories are used as positive samples and the pages of the other categories are used as negative samples. As recommended by the project, we used the *leave-one-university-out* cross-validation method on the 4 universities and the pages of the other universities were always used as training data. A feature set composed by $F_{plain,binary}(G_{cur})$ only was used as the baseline. The results of well performing feature sets compared with prior works are shown in Table 3. The overall results show that tagged-text-based features are consistently effective and the differences caused by the value types are negligible. o-i-e-1_tag and o-i-e-1_tag_real performed the best.

Table 2: Performance of well performing feature sets.

Feature set	Best F-measure	Recall at 99% precision	Precision at 95% recall
o-i-e-1_tag_real	**.811**	*.209*	**.413**
u-1_tag	*.799*	**.232**	.332
all_merged_tag	.775	.179	*.359*
baseline	.748	.180	.265

The results show that our method outperformed all the seven prior works based on macro-averaged F-measure of all the four categories (Macro(4)) and is a little inferior to only one prior work on macro-averaged F-measure of the *course*, *faculty*, and *student* categories (Macro(3)). Our method outperformed 9 out of 12 on a per-category basis (F-measures of the individual categories are not available for four of the previous studies).

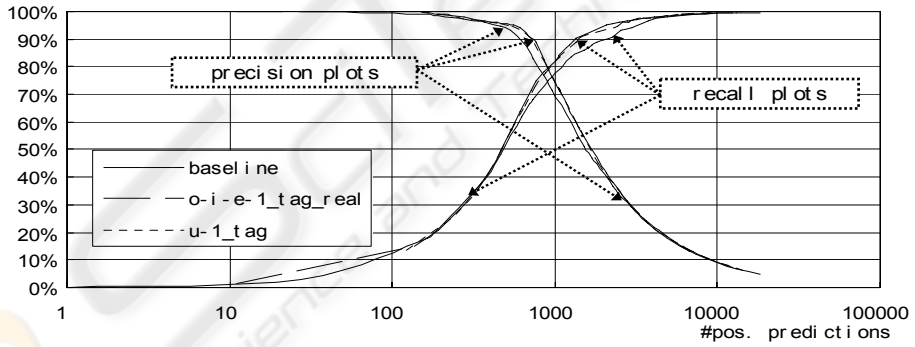

Figure 2: Performance on ResJ-01 data set.

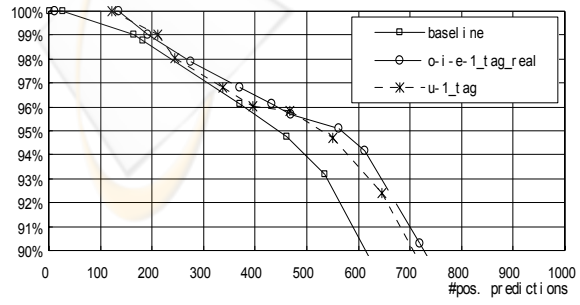

Figure 3: Precision in high precision range.

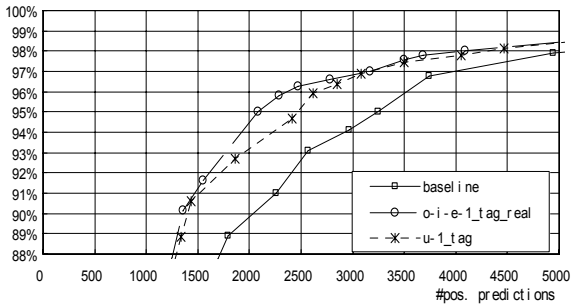

Figure 4: Recall in high recall range.

Table 3: Classification performance on WebKB data set (in percentage of F-measure).

Method		*course*	*faculty*	*project*	*student*	Macro(4)	Macro(3)
Proposed methods	baseline	.684	.760	.400	**.750**	.649	.731
	all_merged	.731	.791	.365	.734	.655	.752
	all_merged_tag	**.802**	**.817**	.393	.685	.674	.768
	u-1_tag	.751	.774	.415	.715	.664	.747
	o-i-e-1_tag	*.778*	*.796*	**.595**	.745	**.729**	**.773**
	o-i-e-1_tag_real	.771	.794	*.572*	**.750**	*.722*	*.772*
Prior works	FOIL(Linked Names) (Yang et al., 2002)					*.629*	
	FOIL(Tagged Words) (Yang et al., 2002)					.591	
	SVM(TA) (Sun et al., 2002)	**.682**	*.659*	*.325*	*.730*	.599	.690
	SVM-FST(XATU) (Kan, 2004)	*.609*	.409	**.665**	.253	.484	.424
	ME(TU) (Kan and Thi, 2005)					.627	
	SVM-iWUM(α =1) (Sun and Lim, 2003)	.547	**.876**	.171	**.958**	**.638**	**.794**
	GE-CKO(FC5) (Sun et al., 2004)						*.765*

Note: The best and the second best results among the proposed methods and among the prior works in each row are shown with bold face and italic face respectively. The prior works that outperformed either o-i-e-1_tag or o-i-e-1_tag_real are shown with underline.

6 CONSIDERATIONS

6.1 Effectiveness of Proposed Features

The experiment results for both the ResJ-01 data set and the WebKB data set indicate that the proposed features are effective on improving the basic performance (in F-measure). In our work, we use the features on page contents of not only the target pages (the baseline) but also the surrounding pages of the target pages. In addition, when using the surrounding pages, the features on different surrounding page groups are not simply merged together (all_merged), but are concatenated so that the context of the surrounding pages can be expressed. That is probably the reason why our method could successfully exploit the surrounding pages while the prior works were not so successful.

Our experiment results for WebKB data set shown in Table 3 are not strictly comparable to the results of the prior works because, although the data set is the same, they used different parts of data or different validation methods. However, the proposed method apparently improved the performance to the highest level of the prior works. More important is that the performance of the proposed method is stable even when the training data size is small.

The proposed feature set is simple and easy to generate and is potentially applicable to wide range of web page classification problems together with various existing classification techniques.

The experiment results for ResJ-01 data set show that the proposed features improve the performance of the precision/recall-assured classifiers as well as the basic performance. The surrounding page groups are consistently effective, but their contributions vary.

For precision-assured classifiers, the upper directory page groups ($G_{*,upper}$) contribute the most to the recall. This probably indicates that such pages provide contextual information, e.g., organization names and research fields, which is lacking in the current page itself but is very important for classifying them with very high confidence.

For recall-assured classifiers, all surrounding page groups ($G_{*,*}$) notably contribute to the precision. This probably indicates that the information from all kinds of surrounding page groups is needed for achieving the high recall, and that the contexts corresponding to the surrounding page structure are appropriately expressed by concatenating the features and successfully reduced the noise which is otherwise introduced by the surrounding pages.

6.2 Reduction of Manual Assessment

To assess how our method would reduce the amount of manual assessment (i.e., the page numbers of *uncertain* output), we compared two compositions of the three-way classifier, one using the baseline and the other using the o-i-e-1_tag_real for both the recall-assured and precision-assured classifiers. Table 4 shows the estimated page numbers of classification output from 'NW100G-01', a 100GB web page corpus for three different quality requirements. Comparing the *uncertain* class sizes, the o-i-e-1_tag_real significantly reduced the amount of pages requiring manual assessment, especially when the required quality is relaxed.

Table 4: Estimated page numbers of classification output from a 100GB web page corpus.

Required quality (precision / recall)	baseline			o-i-e-1_tag_real			Reduction ratio (N_p/N_b)
	assured positive	*uncertain* (N_b)	*assured negative*	*assured positive*	*uncertain* (N_p)	*assured negative*	
99.5% / 98%	3,800	461,832	1,618,988	9,206	358,207	1,717,187	77.6%
99% / 95%	6,163	274,524	1,803,913	11,251	156,782	1,916,567	57.1%
98% / 90%	11,116	155,418	1,918,066	15,503	81,157	1,987,940	52.2%

7 CONCLUSION

We proposed a web page classification method for building a high quality homepage collection, by using a three-way classifier composed of recall-assured and precision-assured component classifiers. By using the proposed method, not only the basic performance in F-measure but also the performance of precision/recall-assured classifiers are improved evidently. At the same time, the effectiveness for reducing the number of pages that need manual assessment to satisfy a required quality given is also shown.

The current classification performance is still far below what can be achieved manually. In the future, we will study a technique to estimate the likelihood of the surrounding pages to be the component pages and incorporate it to the current method.

We also need to investigate the processing cost problem because high performance classifiers require rather complex feature processing that might make it impractical to deal with the enormous size of the web. We have partly tackled this problem with the rough filtering presented in the reference (Wang and Oyama, 2006), and we expect to be able to overcome it by extending the approach.

ACKNOWLEDGEMENTS

This study was partially supported by a Grant-in-Aid for Scientific Research B (No. 18300037) from the Japan Society for the Promotion of Science (JSPS). We used the NW100G-01 document data set with permission from the National Institute of Informatics (NII). We would like to thank Professors Akiko Aizawa and Atsuhiro Takasu of NII for their precious advice.

REFERENCES

Aizawa, A., Oyama, K., 2005. A fast linkage detection scheme for multi-source information integration. In *International Workshop on Challenges in Web Information Retrieval and Integration (WIRI 2005)*, Tokyo, Japan.

CiNii. http://ci.nii.ac.jp/.

Glover, E. J., Tsioutsiouliklis, K., Lawrence, S., Pennock, D. M., Flake, G. W., 2002. Using web structure for classifying and describing web pages. In *11th International World Wide Web Conference*, Honolulu, Hawaii, USA.

Cover, T.M., Thomas, J. A., 1991. *Elements of information theory*. Willey press.

Kan., M.-Y., 2004. Web page categorization without the web page. In *13th World Wide Web Conference (WWW2004)*, New York, NY, USA, May 17-22.

Kan, M.-Y., Thi, H.O.N., 2005. Fast webpage classification using URL features. In *CIKM'05*, Bremen, Germany.

Masada, T., Takasu, A., Adachi, J., 2005. Improving web search performance with hyperlink information. *IPSJ Transactions on Databases*, Vol.46, No.8.

Sun, A., Lim, E.-P., Ng. W.-K., 2002. Web classification using support vector machine. In *4th international workshop on web information and data management.* ACM Press, McLean, Virginia, USA.

Sun, A., Lim, E.-P., 2003. Web unit mining: finding and classifying subgraphs of web pages. In *International Conference on Information and Knowledge Management (CIKM2003)*, New Orleans, Louisiana, USA.

Sun, J., Zhang, B., Chen, Z., Lu, Y., Shi, C., Ma, W., 2004. GE-CKO: A method to optimize composite kernels for web page classification. In *2004 IEEE/WIC/ACM International Conference on Web Intelligence (WI2004)*, Beijing, China.

Wang, Y., Oyama, K., 2006. Combining page group structure and content for roughly filtering researchers' homepages with high recall. *IPSJ Transactions on Databases*, Vol.47, No. SIG 8 (TOD 30).

Yang, Y., Slattery, S., Ghani, R., 2002. A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems*, volume 18. Kluwer Academic Press.