

Feature Selection for a Centroid-Based Websites Classification Approach

Nguyen Viet Thanh, Nguyen Van Khiet

Abstract—Directory service is a useful aid human in looking for information on World Wide Web. A directory service is a pre-categorized list of topics containing many links for each topic. However, most directory services are maintained manually now and face many drawbacks. Therefore, the task of automatic classification of new websites into the topics of directory services becomes very necessary. This paper suggests a new feature selection scheme for the centroid-based websites classification. This scheme gives other weight for terms of the title, metatags and body text of a web page. Our experimental evaluation shows the new classification outperforms the naïve centroid-based classification.

Index Terms—Centroid-based classification, directory service, feature selection, feature vectors, websites classification.

I. INTRODUCTION

INTERNET is a huge source with a lot of information valuable for human. However, people need aid to retrieve this information due to the vast amount and spread distribution. Two common kinds of services available on World Wide Web now are search engines like Google[1] and directory services like Yahoo![2]. While search engine is a website used to find other websites on a particular topic, directory service is a web database best used when searching general topics.

A directory service is a pre-categorized list of topics (subjects) containing many links for each topic. These links may direct to a specific web page but most of them link to a website. We cite this definition from Nassau Library System[3]: A website is a group of similar web pages linked by hyperlinks and managed by a single company, organization, or individual. Websites can range in size from as few as one page to a vast number of pages. Thus, the information presented on a website might belong to various subtopics but follows a common purpose.

Yahoo![2] and most other commercial directory services are still set up manually. We will examine a symbolic directory service, Yahoo!. First, the owner of the website must choose an appropriate category or suggest a new one. Then he or she needs to submit the website to Yahoo!. Finally, the Yahoo!

editors will send an email to announce that Yahoo! accept or reject that registration.

This leads to some drawbacks. First, it will be very time-consuming to extend a large directory service. Second, the links for topics will be soon out of date. Third, a manual directory service will be hard to satisfy some specific user. Fourth, the consistency of classification is hard to maintain since different human experiences are involved. Clearly, we can overcome these drawbacks by using an automatically classifier.

Although text-only classifiers are well-researched and some methods are applied to classify hypertext documents such as web pages, the classification of the whole websites is still being investigated. In this paper, we introduce a feature selection scheme that will improve the websites classification task. We select the title, metatags and body text as features of a web page and assign other weight for terms of them. Each website will be represent by a set of feature vectors and each topic will be represent by a centroid-based feature vector. The new websites will be classified based on how closely its features matches the features of the current websites belonged to different topics,

The organization of this paper is as follows. Section 2 contains a summary of related work in classification of text-only documents, web pages and websites. Section 3 suggests our feature selection scheme. Section 4 describes centroid-based websites classification. Section 5 shows some experimental results. Finally, section 6 provides directions for future research.

II. RELATED WORK

The various text classification algorithms are well research for many years and fall under three general categories. The first category contains algorithms using retrieval techniques such as prototype-based classifier (Rocchio), k-nearest neighbor (KNN), centroid-based classifier [4][5]. The second category includes discriminative classifiers originally from machine learning such as decision-tree, neural networks, support vector machines (SVM). The final category contains generative classifiers such as naïve Bayes classifier. A comparison between these methods is available in [6]. Unfortunately, experimental results in [7] show that many of the above methods perform poorly on hypertext (Yahoo! documents).

Recently, several papers have proposed many methods to

Manuscript received November 17, 2004.

N. V. Thanh and N. V. Khiet are with the Faculty of Information Technology, University of Natural Sciences, Vietnam National University, Hochiminh city, Vietnam (e-mail: nvthanh@fit.hcmuns.edu.vn; e-mail: nvkhiet@fit.hcmuns.edu.vn).

classify hypertext documents. [7] uses several relational learning methods considering existence of links to web pages. [8] considers class labels and the text of neighboring (linked) web pages. [9] applies web page summarization to web page classification. [12] introduces a new tree structure for the hierarchical web documents classification. However, these methods focus on classify single web page, not whole websites.

Some approaches of classifying websites are introduced in [10]. They based on different representations of websites. First approach, a website is represented as a single virtual web page consisting of the union of all its pages. Second approach, a website is represented by a vector of topic frequencies. Final approach, a website is represented by a tree of pages with topics. In [15] a hybrid approach using both local text and hyperlink representation was stated. In [11], a website is represented by a set of feature vectors using term frequency and the authors use a centroid-based classification approach that has satisfactory results. However, this potential approach can be improved by our feature selection scheme based on some specific features of web pages, the title and metatags.

III. A FEATURE SELECTION SCHEME

Feature selection is a major problem related to document classification. In text-only classification, the term weighting scheme is commonly based on the term frequency tf or the combination of term frequency and inverted document frequency $tf-idf$. The term frequency tf_{ij} is the number of occurrences of terms T_j in documents D_j . The inverted document frequency idf_j is the number of documents in which term T_j occurs. From experience, a website contains a set of related meaning web pages with some repeated key terms not rare terms. Therefore the tf_{ij} will plays a much more important role than idf_j in websites classification. So, in our scheme, we just use tf_{ij} representation.

Other much more important features of a HTML document we should consider are the title tag and set of metatags of keywords and descriptions of a web page. These features are used by several major search engines to rank and display results. Naturally, it plays an important role in websites classification. Unfortunately, according to a statistic in [14], not all web pages have metatags and/or quality title. In our set of web pages, this percentage is about one third. So we still have to use body text in website classification.

Based on the above reason we suggest a feature selection based on tf_{ij} on title, metatags and body text with different weight for each category. We call it *weighted term frequency* wtf_{cij} (c denotes *category*). It is formally defined by the following functions.

$$wtf_{cij} = w_c \cdot t_{ij}$$

$$wtf_{ij} = \sum_c wtf_{cij} = \sum_c w_c \cdot t_{ij}$$

We need to estimate w_c for each category. In our experiment

$$\text{we use } w_c = \begin{cases} 3 & \text{if } c = \text{META} \\ 2 & \text{if } c = \text{TITLE} \\ 1 & \text{if } c = \text{BODY} \end{cases}$$

The above definition of website from [3] is too general, so it is difficult to exactly determine whether a web page belonged to a specific website or not. We make a new restricted definition of websites. We limit a website contains only web pages belonging to a unique domain (hostname). For example, we consider www.yahoo.com and dir.yahoo.com belonged to two different domains. With this new definition of website, we can easily identify hyperlinks to web pages not belonged to the website such as advertisement or external links. For instance, while exploring the website, ex. www.yahoo.com, if we find a hyperlink to a new web page outside the website's domain, ex. www.cnn.com/tech, we will not download it. This definition holds for most real world websites and helps us easier to identify a set of web pages representing each website.

Moreover, with the idea that title and metatags text have much more information than body text, we also suggest a breadth first traversal to retrieve web pages of a website. The efficiency of website classification depends on the number of web pages download for each website. However, it is impractical to download all pages so we need to use a heuristic approach to download the most meaningful web pages of a website. My web crawlers will start from the start page and try to extract useful term from the title and metatags. The crawlers will follow the hyperlinks containing these term and again look for the title and metatags content in these pages. If these don't exist, it will choose a hyperlink randomly. We will stop when the number of downloaded web pages is approximately from 100 to 120. For each retrieved page, we perform stop-word removal, word stemming and then build a representative feature vector wtf_{ij} .

Applying this feature selection scheme into the centroid-based classification approach described in Section 4, we can improve the classification performance. We will examine some experimental results in the section 5.

IV. CENTROID-BASED CLASSIFICATION OF WEBSITES

In [4], the authors show that the centroid of text documents is a useful representative of a complete class in terms of KNN classification. We take up this idea and apply it to website classification. The idea of the centroid-based classification is that each topic (website *class*¹) contains several groups of web pages that are somehow related and can be summarized by a common representative, a centroid vector.

Generally, given some groups of related elements, a mean vector for the elements of each group is calculated. We call it centroid vector. A class containing some groups of related elements is represented by the centroid set as the set of all

¹ We shall use the terms (website) *topics* and *class* interchangeably

such centroid vectors. We cite the equation from [11] with a little notation change. Let S be a set of groups of elements e_i with vectors v_{j,e_i} . Let $\pi_g(e_i) = \{v \mid f(v) = g \ \forall v \in e_i\}$ be the restriction of e_i to group g where f is a mapping from a vector v to a group $g \in G$, the set of all groups. Then the centroid set CS of S is defined as:

$$CS(S) = \left[c_j \mid \forall j \in G, c_j = \frac{1}{\left| \bigcup_{\forall i} \pi_g(e_i) \right|} \cdot \sum_{x \in \bigcup_{\forall i} \pi_g(e_i)} x \right]$$

When applying this for website classification, we represent the content of a single page p by a feature vector and so a whole website W by a set of feature vectors.

To get an illustrative view, please refer to Figure 1 showing the idea to represent a sample website class with a centroid set.

We now have two remaining problems. First, we need to choose an appropriate distance measure function. Second, we have to group the similar pages within a website class, a clustering task.

In [4], the authors use a simple cosine function to measure the similarity between a document and a centroid vector. However, in the context of centroid-based website classification, Half Sum of Minimum Distances (HSMD) seems to be the most adequate distance measure between test websites and the centroid sets. For the detail explanation, please refer to [11]. Let W be a website, let C be a centroid set and let $f: P \rightarrow N^d$ be a mapping from P , set of all pages and centroid sets, that returns the feature vector of $p \in P$. The $d(x,y)$, the classic Manhattan function, is used to measure the similarity between two feature vectors due to their much simpler mathematical operations.

$$d(x,y) = \sum_{i=1}^n \left| \frac{x_i}{N_x} - \frac{y_i}{N_y} \right|$$

$$HSMD(W,C) = \frac{\sum_{w_i \in W} \min_{c_j \in C} d(f(w_i), f(c_j))}{|W|}$$

Clustering is well studied for many years with two main approaches, similarity-based and model-based. However, we have to consider some requirements when choosing a clustering algorithm. First, we don't know exactly the number of website topics, so we can eliminate clustering approaches that require inputting the number of clusters (such as K-Means). Second, the chosen clustering algorithm must deal with noise since, in many cases, several pages uncommon for the topic of website they belong to. Third, we don't have any pre training set of websites. Fourth, a major issue is that directory service requires high rate of update in a dynamic environment so the clustering method must adapt it. Considering all above requirements, we choose *GDBSCAN* (Generalized Density-Based Spatial Clustering of Applications with Noise) to group training pages within each website class. For more detail about *GDBSCAN* algorithm

please refer to [13].

So we summarize the algorithm to determine the centroid set for a website class C_i :

1. Collect all web pages (their feature vectors) of the test website of class C_i into one super set.
2. Determine clusters using *GDBSCAN* based on the set of feature vectors.
3. For each cluster, calculate the centroid vector and insert it into the centroid set of class C_i .



Figure 1. Centroid set of a sample website class

V. EXPERIMENTAL EVALUATION

At this time, there is not a standard data set of web pages for websites classification tasks. Thus, like many other papers, we create our own based on Yahoo![2] topics. We just selected 6 topics (of 14 main Yahoo! subject headings) to avoid a too larger data set. For each topic we randomly choose from 10 to 50 example web sites. We also choose 50 web sites from the other topics to make noise. Building a web crawler with breadth first traversal described in Section 3, we retrieve from 100 to 120 for each website. Finally, my data set has 127 web sites containing about 14,000 web pages.

We perform two experiments. First experiment is a binary classification for each topic. The second is a 6-class classification for all topics. The judgment based on precision and recall measurement [16]. We compare between the naive centroid-based classification in [11] (Cent.) with the centroid-based classification using our *weighted term frequency* scheme (Cent.WTF).

For first experiment, there is not big improvement when using Cent.WTF. The results are displayed in Table 1. However, in news topic, Cent.WTF increases the precision considerably. We should pay attention that the Cent. performs weakly in this topics.

Topic	Centroid.		Centroid WTF.	
	pre.	rec.	pre.	rec.
Business	0.75	0.82	0.75	0.82
Computer	0.85	0.71	0.85	0.72
News	0.65	0.88	0.72	0.85
Sports	0.86	0.76	0.87	0.79
Education	0.77	0.82	0.75	0.84
Science	0.79	0.73	0.76	0.75

Table 1. Comparison of Cent. and Cent.WTF in binary classification with precision and recall measurement.

For second experiment, to have a general view of the 6-class classification performance, we just calculate the overall accuracy which is the percentage of correctly classified instances with respect to all tested instances. As the results in Table 2, the Cent.WTF gives a better classification.

	Centroid.	Centroid WTF.
	accuracy	accuracy
6 classes	0.71	0.75

Table 2. Comparison of Cent. and Cent.WTF in 6-class classification with accuracy measurement.

To sum up, the centroid-based classifier using our feature selection scheme outperforms the naïve centroid-based classifier.

VI. CONCLUSION AND FUTURE WORKS

In this paper, we suggest a feature selection scheme to improve the performance of the centroid-based websites classification. This will help in automatically maintaining commercial directory services. The experimental evaluation shows that this is a potential approach. In our future work, we will focus in three tasks.

First, it is clear that not all features contributed equally in distinguishing topics. So we need a scheme to adjust the feature weight. Moreover, when some websites was inserted into the topics, the centroid will change as a result and may be the feature weight need to re-adjust. So we will focus to build an iterative feature weight algorithm.

Second, the centroid-based websites classification now is term-based approach. However, the semantic links of web pages belonged to a website also have an important meaning. We will propose a hybrid approach that uses both term-based and hyperlink-based methods when performing clustering groups of related web pages of a website.

Third, at this time, the centroid-based websites classification just sort websites into topics of the same rank. We have to modify it to adapt hierarchical topics like the topics of commercial directory services.

REFERENCES

- [1] Google: search engine. (<http://www.google.com>)
- [2] Yahoo!: directory service. (<http://dir.yahoo.com>)
- [3] Nassau Library System. (<http://www.nassaulibrary.org>)
- [4] E.H.Han, G.Karypis, "Centroid-based Document Classification: Analysis and Experimental Results", *Proc. 4th PKDD 00*, Lyon, France, 2000.
- [5] S.Shankar, G.Karypis, "Weight adjustment schemes for a centroid-based classifier", *Text Mining Workshop KDD*, 2000.
- [6] Y.Yang, X.Liu, "A re-examination of text categorization methods", *Proc. of the 22nd annual international ACM SIGIR*, 1999.
- [7] S.Chakrabarti, B.Dom, P.Indyk, "Enhanced hypertext categorization using hyperlinks", *Proc. 17th ACM SIMOD*, New York, US, 1998.
- [8] M.Craven, D.DiPasquo, D.Freitag, A.McCallum, T.Mitchell, K.Nigram, S.Slattery, "Learning to Construct Knowledge Bases from the World Wide Web", *Artificial Intelligence*, Elsevier, 1999.
- [9] D.Shen, Z.Chen, Q.Yang, H-J Zeng, B.Zhang, Y. Lu, W-Y.Ma, "Web-page Classification through Summarization", *Proc. ACM SIGIR '04*, Sheffield, UK, 2004.
- [10] M. Ester, H-P. Kriegel, M Schubert, "Website Mining: A new way to spot Competitors, Customers, and Suppliers in the World Wide Web", *Proc. 8th ACM SIGKDD 02*, Alberta, CA, 2002.
- [11] H-P. Kriegel, M Schubert, "Classification of Websites as Sets of Feature Vectors", *Proc. of the LASTED International Conference*, Austria, Feb. 2004.

- [12] W.Wong, A.W.Fu, "Incremental Document Clustering for Web Page Classification", *Proc. of 2000 International Conference on Information Society in the 21st Century: Emerging Technologies and New Challenges (IS2000)*, Japan, 2000.
- [13] J.Sander, M.Ester, H.P.Kriegel, X.Xu "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications", *Data Mining and Knowledge Discovery*, 1998.
- [14] J.M.Pierre, "On the Automated Classification of Web Sites", *Linköping University Electronic Press*, Sweden, 2001.
- [15] D.Riboni, "Feature Selection for Web Page Classification", *EURASIA-ICT 2002 Proceedings of the Workshops*, 2002.
- [16] C.J.V.Rijsbergen, "Information Retrieval", 2nd Edition, 1979.