Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 1735–1747
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Noise reduction through summarization for Web-page classification

Dou Shen [a,*], Qiang Yang [a], Zheng Chen [b]

[a] Department of Computer Science and Technology, Hong Kong University of Science and Technology, Hong Kong, PR China
[b] Microsoft Research Asia, Beijing, PR China

Received 17 July 2006; received in revised form 4 January 2007; accepted 8 January 2007
Available online 26 March 2007

Abstract

Due to a large variety of noisy information embedded in Web pages, Web-page classification is much more difficult than pure-text classification. In this paper, we propose to improve the Web-page classification performance by removing the noise through summarization techniques. We first give empirical evidence that ideal Web-page summaries generated by human editors can indeed improve the performance of Web-page classification algorithms. We then put forward a new Web-page summarization algorithm based on Web-page layout and evaluate it along with several other state-of-the-art text summarization algorithms on the LookSmart Web directory. Experimental results show that the classification algorithms (NB or SVM) augmented by any summarization approach can achieve an improvement by more than 5.0% as compared to pure-text-based classification algorithms. We further introduce an ensemble method to combine the different summarization algorithms. The ensemble summarization method achieves more than 12.0% improvement over pure-text based methods.

Keywords: Web-page categorization; Web-page summarization; Content Body; Noise reduction

1. Introduction

With the rapid growth of the World Wide Web (WWW), there is an increasing need to provide automated assistance to Web users for Web page classification and categorization. Such an assistance is helpful in organizing the vast amount of information returned by keyword-based search engines, or in constructing catalogues that organize Web documents into hierarchical collections; examples of the latter include the Yahoo[1] directory and the LookSmart directory.[2] There is evidence that categorization is expected to play an important role in future search services. For example, research conducted by Chen and Dumais shows that

* Corresponding author.
E-mail addresses: dshen@cse.ust.hk (D. Shen), qyang@cse.ust.hk (Q. Yang), zhengc@microsoft.com (Z. Chen).
[1] http://www.yahoo.com.
[2] http://search.looksmart.com.

0306-4573/$ - see front matter © 2007 Published by Elsevier Ltd.
doi:10.1016/j.ipm.2007.01.013

users prefer navigating through catalogues of pre-classified content (Chen & Dumais, 2000). Such a strong need, however, is difficult to meet without automated Web-page classification techniques due to the labor-intensive nature of human editing.[3]

On a first glance, Web-page classification can borrow directly from the machine learning literature for text classification (Lam & Han, 2003; McCallum & Nigam, 1998; Sebastiani, 2002). On closer examination, however, the solution is far from being so straightforward. Web pages have their own underlying embedded structure in the HTML language. Apart from the main content, they typically contain some noise (for Web-page classification) such as navigation panels, copyright and privacy notices, and advertisements for business purposes and for easy user access. If a pure-text classification method is directly applied to these pages, it will incur much bias for the classification algorithm, making it possible to lose focus on the main topics and important content. Thus, a critical issue is to design an intelligent preprocessing technique to extract the main topic of a Web page.

In this paper, we show that using Web-page summarization techniques for preprocessing in Web-page classification is a viable and effective technique. We further show that instead of using an off-the-shelf summarization technique that is designed for pure-text summarization, it is possible to design specialized summarization methods catering to Web-page structures. In order to collect the empirical evidence that summarization techniques can benefit Web classification, we first conduct an ideal-case experiment, in which each Web page is substituted by its summary generated by human editors. Compared to using the full-text of the Web pages, we gain an impressive 13.2% improvement in F1 measurement. In addition, in this paper, we also propose a new automatic Web summarization algorithm, which extracts the main topic of a Web page by a page-layout analysis to enhance the accuracy of classification. We evaluate the classification performance with this algorithm and compare to some traditional state-of-the-art automatic text summarization algorithms including supervised methods and unsupervised learning methods. Experiment results on LookSmart Web directory show that all summarization methods can improve the micro F1 measure. Finally, we show that an ensemble of summarization methods can achieve more than 12.0% improvement relatively on micro F1 measure, which is very close to the upper bound achieved in our ideal-case experiment.

The rest of the paper is organized as follows. In Section 2, we present the related works on Web classification and summarization. Then we present our proposed unsupervised and supervised summarization algorithms in Section 3. In Section 4, the experimental results on LookSmart Web directory are shown as well as some discussions. Finally, we conclude our work in Section 5.

2. Related work

Recently much work has been done on Web-page summarization (Berger & Mittal, 2000; Buyukkokten, Garcia-Molina, & Paepcke, 2001; Delort, Bouchon-Meunier, & Rifqi, 2003; Sun et al., 2005). Ocelot is a system for summarizing Web pages using probabilistic models to generate the "gist" of a Web page (Berger & Mittal, 2000). The models used are automatically obtained from a collection of human-summarized Web pages. In Buyukkokten et al. (2001), Buyukkokten et al. introduces five methods for summarizing parts of Web pages on handheld devices where the core algorithm is to compute the words' importance using TF/IDF measures and to select important sentences using Luhn's classical method (Luhn, 1958). In Delort et al. (2003), Delort exploits the effect of context in Web-page summarization, which consists of the information extracted from the content of all the documents linking to a page. It is shown that summaries that take into account of the context information are usually more relevant than those made only from the target document. In Sun et al. (2005), Sun et al. extract the knowledge from the clickthrough data to improve the performance of Web-page summarization.

Some research has been conducted to enhance document categorization by summarization (Ker & Chen, 2000; Ko, Park, & Seo, 2004; Kolcz, Prabakarmurthi, & Kalita, 2001), but these works handle pure text categorization only. In Kolcz et al. (2001), Kolcz et al. use summarization as a feature selection method and applies a simple extraction-based technique with several heuristic rules. However, most of these heuristic rules

[3] This article is an extension of Shen et al. (2004).

are hard to apply on Web-page summarization. In Ko et al. (2004), Ko et al. adjust the weights of terms according the importance of the sentence in which the terms appear. The importance of a sentence is determined by two straightforward approaches such as their similarity to the title of the document and the importance of the words contained in the sentence. Similar to Ko et al. (2004), Ker and Chen calculate the weights of a feature(term) in a certain category by borrowing some heuristics from the document summarization literature. Then the predicted categories of a given document are ranked according to the summation of the weights of the features in the document along each category (Ker & Chen, 2000). In this paper, we focus on leveraging state-of-the-art summarization algorithms to remove the noise in Web pages for improving the Web-page classification performance. To cater for the characteristics of Web pages, a new summarization algorithm is proposed and some features related to Web pages are integrated to some existing summarization methods.

In the aspect of removing noise from a Web page, our work is related to Yi, Liu, and Li (2003). Yi et al. propose an algorithm by introducing a tree structure, called Style Tree, to capture the common presentation styles and the actual contents of the pages in a given Web site. However, the Style Tree built on one site is hard to adapt to other sites and it is difficult to build the Style Trees when the number of Web sites is large.

The structure of a Web page is influenced by many factors. Chen, S. Liu, W. Liu, Pu, and Ma (2003) pointed out that when authoring a Web site, the editors usually first conceive the information structure of the site in their mind. They then compile their thoughts into cross-linked Web pages in the HTML language and finally, some extra information, such as navigation bar, advertisement, and copyright information are inserted into the whole page. Since HTML is a visual representation language, much useful information about the content organization is lost after the authoring step. In order to find the important structural information again, two methods have been widely used. One is to extract title and meta-data included in HTML tags to represent the semantic meaning of the Web pages. It is usually true that title and meta-data should be good information to be used by their authors to indicate the main content of Web pages. However, we cannot fully rely on them for the following reasons. First, title and meta-data may be empty in some Web pages. For example, in our data set, about 24.6% of the pages are without any meta-data and 4.8% pages are without a title. Second, some of titles and meta-data may be meaningless since Web-page designer may simply set them by default, with such useless names as ''page1'', ''page2''. Finally, Web site designers may misuse or even give the wrong title or meta-data fields to cheat search engines in order to boost up their ranking.

Therefore, to get satisfying Web-page classification results, it is critical for us to extract the main topic of a Web page by automatically analyzing their context features, such as the anchor text pointing to a Web page (Attardi, Gullí, & Sebastiani, 1999; Chakrabarti, Dom, & Indyk, 1998; Glover, Tsioutsiouliklis, Lawrence, Pennock, & Flake, 2002). In this direction, Glover et al. provided an analysis of the utility of text in citing documents for classification and proved that anchor text was valuable (Glover et al., 2002). Nevertheless, this should be done with care; Chakrabarti studied the role of hyperlink in hypertext classification and pointed out that a naïve use of terms in the linked neighborhood of a Web page could even degrade the classification performance (Chakrabarti et al., 1998).

To summarize, our aim is to apply Web-page summarization to Web-page classification, rather than using pure-text summarization for the purpose. When summarizing a Web page, we rely on the page itself without leveraging extra knowledge such as clickthrough data and anchor text linking to the page. We leave the integration of extra knowledge to our future work.

3. Web-page summarization

In this section, we consider how to analyze the complex implicit structural information embedded in Web pages, and how to use this information for summarization of Web pages. Our approach is to extract the most relevant content from the Web pages and then pass them on to standard text classification algorithms.

In particular, we will consider six different methods for conducting the Web-page summarization. The first method corresponds to an adaptation of Luhn's summarization technique. The second method corresponds to using Latent Semantic Analysis on Web pages for summarization. The third method corresponds to finding the important content body as a basic summarization component. The fourth one is based on graph analysis. The fifth method looks at summarization as a supervised learning task. Finally, we combine the results of all five summarization methods into an ensemble of summarizer.

3.1. Adapted Luhn's summarization method

We adapt Luhn's method that is designed for text summarization for the purpose of Web-page summarization. Luhn's method, described in Luhn (1958), is a systematic approach to perform summarization which forms the core of the field today. In this extraction-based method, every sentence is assigned with a significance factor, and the sentences with the highest significance factor are selected to form the summary. In order to compute the significance factor of a sentence, we need to build a "significant-words pool" which is defined as those words whose frequency is between high-frequency cutoff and low-frequency cutoff that can be tuned to alter the characteristics of the summarization system. After this is done, the significant factor of a sentence can be computed by Luhn's method as follows: (1) set a limit L for the distance at which any two significant words could be considered as being significantly related. (2) Find out a portion in the sentence that is bracketed by significant words not more than L non-significant words apart. (3) Count the number of significant words contained in the portion and divide the square of this number by the total number of words within the portion. The result is the significant factor related to S.

In order to customize this procedure for Web-pages, we make a modification to Luhn's algorithm. In our Web classification task, the category information of each page is already known in the training data, thus significant-words selection could be processed within each category. In this way, we build significant-words pool for each category by selecting the words with high frequency after removing the stop words in that category and then employing Luhn's method to compute the significant factor.

There are two advantages for this modification. First, the prior knowledge of categories is utilized in summarization. Second, some noisy words which may be relatively frequent in an individual page will be removed through the use of statistics over multiple documents. When summarizing the Web pages in the training set, the significant score of each sentence is calculated according to the significant-words pool corresponding to its category label. For a Web page in the test set, we do not have the category information. In this case, we will calculate the significant factor for each sentence according to different significant-words pools over all categories separately. The significant score for the target sentence will be averaged over all categories and referred to as S_{luhn}. The summary of this page will be formed by the sentences with the highest scores.

3.2. Latent semantic analysis (LSA)

Latent semantics analysis (LSA) has been successfully applied to information retrieval (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990) as well as many other related domains. Its power is derived from its ability to represent terms and related concepts as points in a very high dimensional "semantic space" (Landauer, Foltz, & Laham, 1998). In the text summarization area, Gong and Liu (2001) is one of the works that has successfully applied the LSA to pure text. In this section, we will review how to apply LSA to summarization.

To begin with, LSA is based on singular value decomposition (SVD), a mathematical matrix decomposition technique that is applicable to text corpora as known by people. Given an $m * n$ matrix $A = [A_1, A_2, A_n]$, with each column vector A_i representing the weighted term-frequency vector of sentence i in the document under consideration, the SVD is defined as

$$A = U\Sigma V^{\mathrm{T}} \tag{1}$$

where $U = [u_{ij}]$ is an $m * n$ column-orthonormal matrix whose columns are called left singular vectors; $\Sigma = diag(\sigma_1, \sigma_1, \ldots, \sigma_n)$ is an $n \times n$ diagonal matrix whose diagonal elements are non-negative singular values sorted in descending order. $V = [v_{ij}]$ is an $n \times n$ orthonormal matrix whose columns are called right singular vectors (Press, Teukolsky, Vetterling, & Flannery, 1992).

As noted in Berry, Dumais, and O'Brien (1995), Deerwester et al. (1990), LSA is applicable in summarization for two reasons. First, LSA is capable of capturing and modeling interrelationships among terms by semantically clustering terms and sentences. Second, LSA can capture the salient and recurring word combination pattern in a document which describes a certain topic or concept. In LSA, concepts are represented by one of the singular vectors where the magnitude of the corresponding singular value indicates the importance of this pattern within the document. Any sentence containing this word combination pattern will be projected along this singular vector. The sentence that best represents this pattern will have the largest index value with

this vector. We denote this index value as S_{lsa} and select the sentences with the highest S_{lsa} to form the summary. The pseudo-code of SVD-based summarization method can be found in Gong and Liu (2001).

3.3. Content Body identification by page-layout analysis

The structured character of Web pages makes Web-page summarization different from pure-text summarization. This task is difficult due to a number of "noisy" components on a Web page, such as the navigation bar, advertisement, and copyright information. In order to utilize the structure information of Web pages, we employ a simplified version of the Function-Based Object Model (FOM) as described in Chen, Zhou, Shi, Zhang, and Qiu (2001).

In brief, FOM attempts to understand an authors' intention by identifying the object's function and category. In FOM, objects are classified into a Basic Object (BO), which is the smallest information body that cannot be further divided, or a Composite Object (CO) which is a set of Objects (BO or CO) that perform some functions together. An example of a BO is a jpeg file. In HTML contents, a BO is a non-breakable element within two tags or an embedded object. There is no other tag inside the content of a BO. According to this criterion, it is easy to find out all the BOs inside a Web page. Likewise, COs can be detected by a layout analysis of Web pages. The basic idea is that objects in the same category generally have consistent visual styles so that they are separated by apparent visual boundaries, such as table boundaries, from the objects in other categories.

After detecting all the BOs and COs in a Web page, we could identify the category of each object according to some heuristic rules. Detailed examples of these rules are shown in Chen et al. (2001); here we give an overview only. First, the categories of objects include:

Information Object: this object presents content information.
Navigation Object: this object provides navigation guide.
Interaction Object: this object provides user side interaction.
Decoration Object: this object serves for decoration purpose.
Special Function Object: this object performs special functions such as AD, Logo, Contact, Copyright, Reference, etc.

In order to make use of these objects, from the above types of objects, we define the *Content Body* (CB) of a Web page which consists of the main objects related to the topic of that page; these are the objects that convey important information about the page. The algorithm for detecting CB is as follows:

- Consider each selected object as a single document and build the TF $*$ IDF index for the object.
- Calculate the similarity between any two objects using Cosine similarity computation, and add a *link* between them if their similarity is greater than a threshold. The threshold is chosen empirically. After processing all pairs of objects, we will obtain a linked graph to connect different objects.
- In the graph, a *core object* is defined as the object having the most edges.
- Extract the CB as the combination of all objects that have edges linked to the core object.

Finally, we will assign a score S_{cb} to each sentence, for which $S_{cb} = 1.0$ if the sentence is included in "content body"; otherwise, $S_{cb} = 0$. Finally, all sentences with S_{cb} equal to 1.0 give rise to the summary of the Web page in question.

3.4. Graph-based summarization

A Web page can be represented as a graph in which each vertex is a sentence and the weight of a edge linking two nodes is the similarity between the corresponding sentences. After constructing the graph, some popular graph-based ranking algorithms can be applied to rank the sentences according to certain measurements. Then we can select the top sentences into summary. Mihalcea (2005) studied the application of PageRank and HITS on this problem, where she tried to construct both directed and undirected graph from a document.

Take HITS as an example, the weight of a vertex (corresponding to a sentence), which can be an "authority score" or "hub score" is calculated iteratively by

$$\begin{aligned} \mathrm{HITS_A}(V_i) &= \sum_{V_j \in In(V_i)} \mathrm{HITS_H}(V_j) \\ \mathrm{HITS_H}(V_i) &= \sum_{V_j \in Out(V_i)} \mathrm{HITS_A}(V_j) \end{aligned} \tag{2}$$

Among the several options studied in Mihalcea (2005), the method based on the authority score of HITS on the directed backward graph performs best. Therefore, it is taken as a candidate summarization approach in this paper. We denote this method by "HITS" and the generated score for each sentence by S_{HITS}. More details about the graph-based summarization algorithms can be found in Mihalcea (2005).

3.5. Supervised summarization

Besides the unsupervised summarization algorithms described above, some researchers also focus on generating the summary using machine learning approaches (Berger & Mittal, 2000; Chuang & Yang, 2000; Kupiec, Pedersen, & Chen, 1995; Teufel & Moens, 1997; Yeh, Ke, Yang, & Meng, 2005). In this paper, we also employ a supervised approach for Web summarization, by making full use of the labeled training data. A set of features are first extracted from each of a Web page. Then, a supervised learning algorithm is applied to train the summarizer to identify whether a sentence should be selected into its summary or not. There are a total of eight features utilized in our algorithm, where five of them are common features for text document and Web page and the rest three of them are specific to Web page layout. Some notations are defined as follows:

- PN: the number of paragraphs in a document;
- SN: the number of sentences in a document;
- PL_k: the number of sentences in a certain paragraph k;
- $\mathrm{Para}(i)$: the associated paragraph of sentence i;
- TF_w: the number of occurrences of word w in a target Web page;
- SF_w: the number of sentences including the word w in the Web page;

Given a set of sentences S_i $(i = 1 \ldots \mathrm{SN})$ in a page, the eight features are defined as follows:

- f_{i1} measures the position of a sentence S_i in a certain paragraph.
- f_{i2} measures the length of a sentence S_i, which is the number of words in S_i.
- $f_{i3} = \sum \mathrm{TF}_w * \mathrm{SF}_w$. This feature takes into account not only the number of word w into consideration, but also its distribution among sentences. We use it to punish the locally frequent words.
- f_{i4} is the similarity between S_i and the title. This similarity is calculated as the cosine similarity between the sentence and the title. Both the sentence and the title are represented by TF vectors. The same representation method is used for f_{i5} and f_{i6}.
- f_{i5} is the cosine similarity between S_i and all text in the page.
- f_{i6} is the cosine similarity between S_i and meta-data in the page.
- f_{i7} is the number of occurrences of word from S_i in special word set. The special word set is built by collecting the words in the Web page that are italic or bold or underlined.
- f_{i8} is the average font size of the words in S_i. In general, larger font size in a Web page is given higher importance.

After extracting these eight features from a Web page, we apply the Naïve Bayesian classifier to train a summarizer for simplicity, as in Kupiec et al. (1995):

$$p(s \in S | f_1, f_2, \ldots, f_8) = \frac{\prod_{j=1}^{8} p(f_j | s \in S) p(s \in S)}{\prod_{j=1}^{8} p(f_j)} \tag{3}$$

where $p(s \in S)$ stands for the compression rate of the summarizer, which can be predefined for different applications, $p(f_i)$ is the probability of each feature i and $p(f_i|s \in S)$ is the conditional probability of each feature i. The latter two factors can be estimated from the training corpus. Each sentence will then be assigned a score by the above equation, which is denoted as S_{sup}.

3.6. An ensemble of summarizers

From the previous approaches, we can build various summaries independently for a Web page. These approaches are based on different mechanisms and can be complementary to each other. Therefore, it is beneficial to combine them. Similar problems have been studied widely in the field of machine learning and previous works have shown that a proper combination can make improvement (Alpaydin, 2004; Dietterich, 2000; Hansen & Salamon, 1990). In this paper, we plan to combine the different summarization approaches through a weighted linear combination, which is proved to be effective (Shen et al., 2005):

$$S_{\text{ensemble}} = w_1 S_{\text{luhn}} + w_2 S_{\text{lsa}} + w_3 S_{\text{cb}} + w_4 S_{\text{HITS}} + w_5 S_{\text{sup}} \quad (4)$$

where $w_i (1 = 1, 2, \ldots, 5)$ is the weight for each base summarization method. A detailed study of the weights is presented in the experiments. After combining the scores, the sentences with the highest S_{ensemble} will be chosen into the summary.

4. Experiments

In order to test the effectiveness of summarization for Web classification, several experiments are conducted. Firstly, we test the Web-page classification on the human created summaries in order to find out whether summarization can help classification of Web pages at all. Having confirmed this hypothesis, we compare our proposed "content body identification summarizer" with three unsupervised methods including: adapted Luhn's algorithm, LSA-based method and the graph-based method, as well as the supervised summarizer. Finally, our ensemble of summarizers is evaluated. In our experiments, we also study the variation of different parameter settings for composing the best summarizer.

4.1. Data set

In our experiments, we use about 2 million Web pages crawled from the LookSmart Web directory (http://search.looksmart.com). Due to the limitation of network bandwidth, we only downloaded about 500 thousand descriptions of Web pages that are manually created by human editors. Since it is a time-consuming task to run experiments on this large data set, we randomly sampled 30% of the pages with descriptions for our experiment purpose. The extracted subset includes 153,019 pages, which are distributed among 64 categories (we only consider the top two level categories on LookSmart Website). The largest category (Library\Society) consists of 17,473 pages; while the smallest category (People & Chat\Find People) consists of only 52 pages. Tables 1 and 2 show the number of pages for the three largest categories and three smallest categories. In order to reduce the uncertainty of data split, a 10-fold cross-validation procedure is applied in our experiments.

4.2. Classifiers

Since the focus of this paper is to test the effectiveness of Web summarization for classification, we choose two popular classifiers in our experiments. One is a naïve Bayesian classifier (McCallum & Nigam, 1998; Mitchell, 1997), and another is a support vector machine (Cortes & Vapnik, 1995; Joachims, 1998).

Table 1
The three largest categories

Category name	Total	Train	Test
Library\Society	17,473	15,726	1747
Travel\Destinations	3324	11,992	1332
Entertainment\Celebrities	10,112	9101	1011

Table 2
The three smallest categories

Category Name	Total	Train	Test
Sports\News & Scores	106	96	10
People & Chat\Personals	74	67	7
People & Chat\Find People	52	47	5

4.2.1. Naïve Bayesian classifier (NB)

The Naïve Bayesian classifier (NB) is a simple but effective text classification algorithm which has been shown to perform very well in practice (McCallum & Nigam, 1998; Mitchell, 1997). The basic idea in NB is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document. As described in McCallum and Nigam (1998), most researchers employ NB method by applying Bayes' rule:

$$P(c_j|d_i;\hat{\theta}) = \frac{P(c_j|\hat{\theta})\prod_{k=1}^{n} P(w_k|c_j;\hat{\theta})^{N(w_k,d_i)}}{\sum_{r=1}^{|C|} P(c_r|\hat{\theta})\prod_{k=1}^{n} P(w_k|c_r;\hat{\theta})^{N(w_k,d_i)}} \tag{5}$$

where $P(c_j|\hat{\theta})$ can be calculated by counting the frequency with each category c_j occurring in the training data; $|C|$ is the number of categories; $p(w_i|c_j)$ stands for probability that word w_i occurs in class c_j which maybe small in training data, so the Laplace smoothing is chosen to estimate it; $N(w_k, d_i)$ is the number of occurrences of a word w_k in d_i; n is the number of words in the training data.

4.2.2. Support vector machine (SVM)

Support vector machine (SVM) is a powerful learning method introduced by Vapnik et al. (Cortes & Vapnik, 1995; Joachims, 1998; Vapnik, 1995). It is well founded in terms of computational learning theory and has been successfully applied to text categorization (Joachims, 1998, 1999). SVM operates by finding a hyper-surface in the space of possible inputs. The hyper-surface attempts to split the positive examples from the negative examples by maximizing the distance between the nearest of the positive and negative examples to the hyper-surface. Intuitively, this makes the classification correct for testing data that is near but not identical to the training data. There are various ways to train SVMs. One particularly simple and fast method is Sequential Minimal Optimization (SMO) first described by J. Platt which is available online.[4] His sequential minimal optimization algorithm breaks the large quadratic programming (QP) problem down into a series of small QP problems to be solved analytically. Thus the SMO algorithm is efficiently applicable for large feature and training sets. In this paper, a linear kernel is used and the one-against-rest approach is applied for the multi-class case.

4.3. Evaluation measure

We employ the standard measures to evaluate the performance of Web classification, i.e. precision, recall and F1-measure (van Rijsbergen, 1979). Precision (P) is the proportion of actual positive class members returned by the system among all predicted positive class members returned by the system. Recall (R) is the proportion of predicted positive members among all actual positive class members in the data. $F1$ is the harmonic average of precision and recall as shown below:

$$F1 = 2 \times P \times R/(P + R) \tag{6}$$

To evaluate the average performance across multiple categories, there are two conventional methods: micro-average and macro-average. Micro-average gives equal weight to every document; while macro-average gives equal weight to every category, regardless of its frequency. In our experiments, we used both of them to evaluate the performance of classification. However, we report only the micro-average for simplicity since we came to the same conclusion in terms of macro-average.

[4] http://research.microsoft.com/jplatt/smo.html.

4.4. Experimental results and analysis

4.4.1. Baseline

A simple way to perform Web classification is to treat it as a pure-text document. In our experiment, the state-of-the-art text classification algorithms (NB&SVM) are applied to build the baseline system. Firstly, Web pages are converted to pure text document by removing the HTML tags. Then, each document is tokenized with a stop-word remover and Porter stemming.[5] Finally, each Web page is represented by bag-of-words, in which the weight of each word is assigned with their term frequency. In order to speed-up the classification, a simple feature selection method, ''document frequency selection (DF)'' (Yang & Pedersen, 1997), is applied in our experiment. In our experiments, the words whose DF is lower than a threshold are removed from feature set. The threshold is set as 6 empirically in this paper. Finally, we obtain the classification results based on the selected word features, as shown in the ''Full-text'' row of Tables 3 and 4. From these two tables, we found that SVM achieves 0.651 in micro-F1, which outperform the NB result by about 2.4% relatively. We also found that the variance of 10-fold cross-validation is quite small (about 0.3%), which indicates that the classification is stable on this data set.

4.4.2. Results of human's summary

In order to test the effectiveness of summarization techniques for Web classification, we conduct a feasibility study in our experiment. We extract the description of each Web page from the LookSmart Website and consider it as the ''ideal'' summary for the page. Since the description is authored by the Web directory editors, the quality is considered to be good enough to be the summary for the page. We apply the classifiers directly on these descriptions instead of the full-text of the Web pages. This experiment can help us understand whether in the best case, summarization can help improve the classification. In addition, the title and meta-data of a Web page can also be considered as a kind of summary. An example of the description, title and meta-data is shown in Table 5 and the classification results on these ''ideal summary'' are shown in the related rows of Tables 3 and 4. Compared to full-text classification, classification on human-authored ''description'' can significantly improve the F1 measure by more than 13.2% using either classifier. However, classification on ''pure title'' or ''pure meta-data'' achieves worse $F1$-measure results as compared to the baseline system; this is because ''pure title'' or ''pure meta-data'' is usually short and do not contain sufficient information. Through digging into the data, we found that Web-page ''descriptions'' can easily help the end-user to understand the meaning of the Web page. Although the title can play this role also to some extent, their short lengths indeed make it impossible to represent the full meaning of the page. The uneven quality of the meta-data prevents them from achieving good results because much of the meta-data is assigned by the designers with default values.

Through the ''ideal-case'' experiments, we have found that the ''ideal summary'' (such as ''descriptions'') can indeed help improve the Web classification performance. In addition, if the summary is not done properly, then the ''bad summary'' can hurt the performance. Hence, in the rest of the experiments, we hope to achieve a similar ''good'' summary by our automatic Web summarization techniques.

4.4.3. Results of unsupervised summarization algorithms

In this section, we evaluate the unsupervised summarization methods, including our proposed ''content body identification'' summarizer, adapted Luhn's algorithm, LSA-based methods and the graph-based methods.

As mentioned in Section 3.3, for the ''content body identification'' summarizer, we set a threshold value to determine whether there is a link between the two objects on a Web page. In our experiment, the threshold is set to be 0.1. Through our experiments, we found that most of the unrelated objects in Web pages, such as copyright and advertisement banner, can be easily removed by our algorithm. For the adapted Luhn's algorithm, the LSA algorithm and the graph-based algorithm, the compression rate is set as 20%, 30% and 30%, respectively in our experiments. A detailed study of the above parameters is given in Section 4.4.6. From

[5] http://tartarus.org/martin/PorterStemmer/.

Table 3
Experimental results of NB (%)

	micro$P$	micro$R$	micro-$F1$
Full-text	70.7 ± 0.3	57.7 ± 0.3	63.6 ± 0.3
Title	68.3 ± 0.4	55.4 ± 0.4	61.2 ± 0.4
Meta-data	47.7 ± 0.4	38.7 ± 0.4	42.7 ± 0.4
Description	81.5 ± 0.4	66.2 ± 0.4	73.0 ± 0.4
Content Body	77.2 ± 0.4	62.7 ± 0.4	69.2 ± 0.4
Luhn	77.9 ± 0.4	63.3 ± 0.4	69.8 ± 0.5
LSA	75.9 ± 0.4	61.7 ± 0.4	68.1 ± 0.5
HITS	78.2 ± 0.4	63.5 ± 0.4	70.1 ± 0.4
Supervised	75.2 ± 0.4	60.9 ± 0.4	67.3 ± 0.4
Hybrid	80.5 ± 0.3	65.2 ± 0.3	72.0 ± 0.3

Table 4
Experimental results of SVM (%)

	micro$P$	micro$R$	micro-$F1$
Full-text	72.4 ± 0.3	59.3 ± 0.3	65.1 ± 0.3
Title	68.8 ± 0.3	55.9 ± 0.3	61.7 ± 0.3
Meta-data	47.8 ± 0.4	38.8 ± 0.4	42.8 ± 0.4
Description	82.1 ± 0.4	66.9 ± 0.4	73.7 ± 0.4
Content Body	78.6 ± 0.3	63.7 ± 0.3	70.3 ± 0.3
Luhn	77.3 ± 0.3	62.8 ± 0.3	69.3 ± 0.3
LSA	79.2 ± 0.3	64.3 ± 0.3	71.0 ± 0.3
HITS	79.7 ± 0.4	64.4 ± 0.3	71.2 ± 0.3
Supervised	76.3 ± 0.4	61.8 ± 0.4	68.3 ± 0.4
Hybrid	81.4 ± 0.3	66.1 ± 0.3	72.9 ± 0.3

Table 5
An example of the human-supplied "good summary": the description, title and meta-data of a page

Item	Content
Description	AAP – Do Yourself a Favor: Skip the Tan Warns about the effects of suntans, including wrinkles and skin cancer. From the American Academy of Pediatrics
Title	AAP – Do Your Skin a Favor: Skip the Spring Break Tan
Meta-data	Null

Tables 3 and 4, we found that these four unsupervised summarization algorithms are comparable on classification experiment, although the graph-based method outperforms the other three a little bit with either NB or SVM. We can see that all of them can achieve more than 6.5% improvement as compared to the baseline system, which verifies the effectiveness of the unsupervised summarization methods on reducing noise in Web pages for classification.

4.4.4. Result on supervised summarization algorithm

In order to construct a supervised summarizer as described in Section 3.5, we need some training data consisting of the pairs of a Web page and the corresponding summary generated by human, where a summary is composed of the sentences extracted from a Web page. However, in our data set, the Web-page description is authored by Web-directory editors instead of extracted from the Web pages, we need to tag each sentence as positive or negative example for collecting the training data. In our experiment, we define one sentence as positive if its similarity with the description is greater than a threshold (0.3 in this paper), and others as negative. The $F1$ measure of the supervised method (denoted by Supervised) is shown in Tables 3 and 4 (when compression rate equals to 20%). We found it can achieve about 5.0% relative improvement compared to baseline system, which is a little worse than unsupervised algorithms. The reason may be that our training data selection is

not precise since we only rely on the similarity to descriptions. However, the improvement also verifies the effectiveness of summarization methods for classification as proved by the unsupervised summarization methods.

4.4.5. Result on hybrid summarization algorithm

Through the above experiments, we found that both unsupervised and supervised summarization algorithms can improve the classification accuracy to some extent. But none of them can approach the upper bound of the system set by human edited summary. Since all of them use different mechanisms, their results may be complementary to each other and a proper ensemble of them is beneficial. The experimental results shown in Tables 3 and 4 verify our hypothesis. As we can see, the ensemble of summarization methods can achieve about an impressive 12.0% improvement as compared to baseline system, which is also very near to the upper bound of the system. In this experiment, we use the same weighting for each summarization algorithm. We will consider the different weighting schema in Section 4.4.7.

4.4.6. Performance using different compression rates

In this section, we conduct an experiment to find the relationship between the performance of classification and the compression rate of summarization. We report only the experimental results with NB for simplicity. The results are shown in Tables 6 and 7.

From Tables 6 and 7, we found that all the methods reach their peak performance when the compression rate is 20% or 30% (for "Content Body" when the threshold equals to 0.10). When the compression rate rises to 50%, the performance of some methods such as LSA and supervised summarization method become worse than the baseline. This may be ascribed to the inclusion of noises with the raise of the compression rate. However, in the wide range from 10% to 40%, all the summarization methods can improve the classification performance to some extent over baseline.

4.4.7. Effect of different weighting schemata

In the section, experiments are conducted to test the effect of different weighting schema for the ensemble of the different summarizers, with NB as the classifier. We tested six cases denoted by Schemas1–6 which assigns different weights (as shown in Eq. (4)) to different summarization scores, in addition to the original schema which sets an equal weight for each summarization algorithm.

Schema1–5: We increase the value of w_i $(i = 1, 2, 3, 4, 5)$ to two in Schemas1–5 respectively and keep others as one.
Schema6: We assign the weight of each summarization method in proportion to the performance of each method (the value of micro-$F1$).

From the results shown in Table 8, we can see that different schemata made no obvious difference.

Table 6
Performance of "Content Body" using different thresholds with NB

	0.25	0.20	0.15	0.10	0.05
Content Body	64.1 ± 0.4	65.0 ± 0.5	67.0 ± 0.4	69.2 ± 0.4	66.7 ± 0.3

Table 7
Performance of different compression rates with NB

	10%	20%	30%	40%	50%
Luhn	66.1 ± 0.5	69.8 ± 0.5	67.4 ± 0.4	66.1 ± 0.3	64.5 ± 0.3
LSA	66.3 ± 0.6	67.0 ± 0.5	68.1 ± 0.5	65.9 ± 0.3	63.4 ± 0.3
HITS	67.5 ± 0.5	68.8 ± 0.4	70.1 ± 0.4	67.0 ± 0.3	65.8 ± 0.3
Supervised	66.1 ± 0.5	67.3 ± 0.4	64.8 ± 0.4	64.4 ± 0.3	62.9 ± 0.3
Hybrid	67.2 ± 0.4	69.6 ± 0.4	72.0 ± 0.3	70.1 ± 0.3	67.0 ± 0.3

Table 8
Effect of different weighting schema with NB

	micro$P$	micro$R$	micro-$F1$
Origin	80.5 ± 0.3	65.2 ± 0.3	72.0 ± 0.3
Schema1	81.4 ± 0.4	65.9 ± 0.4	72.8 ± 0.4
Schema2	79.7 ± 0.4	64.5 ± 0.4	71.3 ± 0.4
Schema3	81.1 ± 0.3	65.6 ± 0.3	72.5 ± 0.3
Schema4	81.5 ± 0.3	66.0 ± 0.4	72.9 ± 0.3
Schema5	79.8 ± 0.4	64.9 ± 0.4	71.6 ± 0.4
Schema6	80.7 ± 0.3	65.9 ± 0.3	72.6 ± 0.3

5. Conclusions and future work

In this paper, several Web-page summarization algorithms are proposed for removing the noise from Web pages while keeping the most relevant features for improving the accuracy of Web classification. As illustrated by our ideal-case experiment, the summary created by human editors can achieve more than 13.2% improvement by the micro-$F1$ measure as compared to the pure text of the Web pages. This observation validates the need to find better Web-page summarization methods. We evaluated Web-page categorization on several state-of-the-art automatic document summarization algorithms, as well as an algorithm by utilizing the layout analysis of Web pages. Experimental results show that automatic summary can achieve a similar improvement (above 12.0% improvement) as the ideal-case accuracy achieved by using the summary created by human editors.

In this paper, we only considered a Web page as an isolated document. However, more and more research works demonstrate that the hyperlink is one of the important features for Web search and analysis. In the future, we will investigate methods for multi-document summarization of the hyperlinked Web pages to boost the accuracy of Web classification. Besides that, we hope to leverage more information such as clickthrough data to increase the summarization performance so that we can further improve the classification result.

References

Alpaydin, E. (2004). *Introduction to machine learning*. The MIT Press.

Attardi, G., Gullí, A., & Sebastiani, F. (1999). Automatic web page categorization by link and context analysis. In C. Hutchison, G. Lanzarone (Eds.), *Proceedings of THAI-99, first European symposium on telematics, hypermedia and artificial intelligence* (pp. 105–119). Varese, IT.

Berger, A. L., & Mittal, V. O. (2000). Ocelot: A system for summarizing web pages. In *SIGIR'00: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 144–151).

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review, 37*(4), 573–595.

Buyukkokten, O., Garcia-Molina, H., & Paepcke, A. (2001). Seeing the whole in parts: Text summarization for web browsing on handheld devices. In *WWW'01: Proceedings of the 10th international conference on World Wide Web* (pp. 652–662).

Chakrabarti, S., Dom, B. E., & Indyk, P. (1998). Enhanced hypertext categorization using hyperlinks. In L. M. Haas & A. Tiwary (Eds.), *Proceedings of SIGMOD-98, ACM international conference on management of data* (pp. 307–318). New York, US, Seattle, US: ACM Press.

Chen, H., & Dumais, S. T. (2000). Bringing order to the web: Automatically categorizing search results. In *Proceedings of CHI-00, ACM international conference on human factors in computing systems* (pp. 145–152). New York, US, Den Haag, NL: ACM Press.

Chen, Z., Liu, S., Liu, W., Pu, G., & Ma, W.-Y. (2003). Building a web thesaurus from web link structure. In *SIGIR'03: Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval* (pp. 48–55).

Chen, J., Zhou, B., Shi, J., Zhang, H., & Qiu, F. (2001). Function-based object model towards website adaptation. In *WWW'01: Proceedings of the 10th international conference on World Wide Web* (pp. 587–596).

Chuang, W. T., & Yang, J. (2000). Extracting sentence segments for text summarization: a machine learning approach. In: *SIGIR'00: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 152–159).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS, 41*(6), 391–407.

Delort, J.-Y., Bouchon-Meunier, B., & Rifqi, M. (2003). Web document summarization by context. In *WWW (Posters)*.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *MCS'00: Proceedings of the first international workshop on multiple classifier systems* (pp. 1–15). London, UK: Springer-Verlag.

Glover, E. J., Tsioutsiouliklis, K., Lawrence, S., Pennock, D. M., & Flake, G. W. (2002). Using web structure for classifying and describing web pages. In *Proceedings of WWW-02, international conference on the World Wide Web* (pp. 562–569). New York, US, Honolulu, US: ACM Press.

Gong, Y., & Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. In *SIGIR'01: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 19–25).

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12*(10), 993–1001.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *Proceedings of ECML-98, 10th European conference on machine learning* (pp. 137–142). Heidelberg, DE, Chemnitz, DE: Springer Verlag, published in the "Lecture Notes in Computer Science" series, number 1398.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In I. Bratko & S. Dzeroski (Eds.), *Proceedings of ICML-99, 16th international conference on machine learning* (pp. 200–209). San Francisco, US, Bled, SL: Morgan Kaufman.

Ker, S. J., & Chen, J.-N. (2000). A text categorization based on summarization technique. In *Proceedings of the ACL-2000 workshop on recent advances in natural language processing and information retrieval* (pp. 79–83).

Ko, Y., Park, J., & Seo, J. (2004). Improving text categorization using the importance of sentences. *Information Processing and Management, 40*(1), 65–79.

Kolcz, A., Prabakarmurthi, V., & Kalita, J. K. (2001). String match and text extraction: Summarization as feature selection for text categorization. In H. Paques, L. Liu, & D. Grossman (Eds.), *Proceedings of CIKM-01 10th ACM international conference on information and knowledge management* (pp. 365–370). New York, US, Atlanta, US: ACM Press.

Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. In *SIGIR'95: Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 68–73).

Lam, W., & Han, Y. (2003). Automatic textual document categorization based on generalized instance sets and a metamodel. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*(5), 628–633.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*, 259–284.

Luhn, H. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development, 2*(2).

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. In *Proceedings of AAAI-98, workshop on learning for text categorization*.

Mihalcea, R. (2005). Language independent extractive summarization. In *AAAI* (pp. 1688–1689).

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill Higher Education.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in C++: The art of scientific computing*. New York: Cambridge University Press.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1–47.

Shen, D., Chen, Z., Yang, Q., Zeng, H.-J., Zhang, B., Lu, Y., et al. (2004). Web-page classification through summarization. In *SIGIR'04: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 242–249). New York, NY, USA: ACM Press.

Shen, D., Pan, R., Sun, J.-T., Pan, J. J., Wu, K., Yin, J., et al. (2005). Q2C@UST: Our winning solution to query classification in kddcup 2005. *SIGKDD Explorations, 7*(2), 100–110.

Sun, J.-T., Shen, D., Zeng, H.-J., Yang, Q., Lu, Y., & Chen, Z. (2005). Web-page summarization using clickthrough data. In *SIGIR'05: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 194–201). New York, NY, USA: ACM Press.

Teufel, S., & Moens, M. (1997). A comparison of event models for naive bayes text classification. In *Proceedings of ACL/EACL-97: Workshop on intelligent and scalable text summarization*.

van Rijsbergen, C. (1979). *Information retrieval* (2nd ed.). London: Butterworth.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In D. H. Fisher (Ed.), *Proceedings of ICML-97, 14th international conference on machine learning* (pp. 412–420). San Francisco, US, Nashville, US: Morgan Kaufman Publishers.

Yeh, J.-Y., Ke, H.-R., Yang, W.-P., & Meng, I.-H. (2005). Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing and Management, 41*(1), 75–95.

Yi, L., Liu, B., & Li, X. (2003). Eliminating noisy information in web pages for data mining. In: *KDD'03: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 296–305).