Exploring in the Weblog Space by Detecting Informative and Affective Articles

Xiaochuan Ni[1], Gui-Rong Xue[1], Xiao Ling[1], Yong Yu[1], Qiang Yang[2]

[1]Department of Computer Science and Engineering
Shanghai Jiao-Tong University
Shanghai, P.R.China
+86-21-54745879*608

[2]Department of Computer Science
Hong Kong University of Science and Technology
Clearwater Bay, Kowloon, Hong Kong
+852-23588768

{nixiaochuan, grxue, shawnling, yyu}@apex.sjtu.edu.cn qyang@cs.ust.hk

ABSTRACT

Weblogs have become a prevalent source of information for people to express themselves. In general, there are two genres of contents in weblogs. The first kind is about the webloggers' personal feelings, thoughts or emotions. We call this kind of weblogs affective articles. The second kind of weblogs is about technologies and different kinds of informative news. In this paper, we present a machine learning method for classifying informative and affective articles among weblogs. We consider this problem as a binary classification problem. By using machine learning approaches, we achieve about 92% on information retrieval performance measures including precision, recall and F1. We set up three studies on the applications of above classification approach in both research and industrial fields. The above classification approach is used to improve the performance of classification of emotions from weblog articles. We also develop an intent-driven weblog-search engine based on the classification techniques to improve the satisfaction of Web users. Finally, our approach is applied to search for weblogs with a great deal of informative articles.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous; I.5.4 [**Pattern Recognition**]: Application- *Text processing*

General Terms

Algorithms, Measurement, Performance, Design, Experimentation, Human Factors.

Keywords

Weblog, Informative article, Affective article, Classification, User Intent.

1. INTRODUCTION

Weblogs, also referred to as blogs, are generally considered as online diaries published and maintained by individual users (bloggers), reporting on the bloggers' daily activities and feelings. Compared with traditional media such as online news sources and public websites maintained by companies, blogs mainly have two unique characteristics [19]: (1) they are mainly maintained by individual persons and thus the contents are generally personal, and (2) the link structures between blogs generally form localized communities.

In the recent few years, the number of blogs has increased dramatically. According to a report in [9], by the end of 2005, there were 100 million blogs on the global Internet, of which 16 million were from China, representing 1600 time increase in 4 years. It has been estimated that by the end of 2007 there may be as many as 100 million blogs in China. Using the blogs, people have dramatically changed the way in which they show their feelings and publish news and other information that they are interested in. In September 2005, 27% of the Web users in America had used the Internet to read someone else's blog, and by January 2006, that figure had increased to 39% [22] [23].

The blog fever is accompanied by increasing interests from research and industrial communities to harness this important information source. Much research work is being conducted on blogs. Ongoing research in this area includes content based analysis [1] [4] [5] [19] and blog communities' evolution [14] [15], which focuses on the blogs' different characteristics respectively. As more people begin to write blogs, different kinds of tools are also being provided to help users retrieve, organize and analyze the blogs. The need for blog-search engines and blog tagging systems [7] [11] is also on the rise.

We consider two main genres in blog's content. First, there is the online dairy by which people share their daily life publicly, express their feelings or thoughts or emotions through the blogs. In this paper, this genre is considered as affective. A second kind of blogs is topic-oriented; the topic can be related to a hobby or the author's profession or business [4]. In this paper, this genre of content is called informative. The increase in the amount of weblogs drives users to access textual information in new ways and information needs differ with different types of available information. It is highly desirable to identify the informative and affective contents automatically in blogs. This desire is beneficial to many applications, such as blog search and topic and emotion classification of blog articles, etc.

In this paper, we address the task of separating informative articles from affective articles in blogs. In addition, we also focus our attention on the studies of its applications on promoting the development of other research work and inventing new methods and tools to improve user experience and satisfaction while surfing in the blog space.

In this paper, we consider the genre-detection problem as a binary classification problem, which we solve by using text classification techniques. This problem poses a number of scientific challenges. (1) The definitions of the informative articles and the affective

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

articles should be comprehensive and consistent with the intentions of bloggers and visitors. (2) The training corpus for both categories should be appropriately sampled. (3) The machine learning algorithms should be carefully chosen and adapted to solve this problem with high effectiveness.

Our approach is as follows. We first conduct a user study to collect descriptions of informative articles and affective articles. We define these two genres of articles by summarizing the most consistent descriptions. We select a number of articles from the traditional public website as the training samples for the informative category. Because these articles are well formed and written, and have been published through editing process, we consider them to be of high quality. We also ask some human labelers to tag blog articles using the two categories. For the labeled articles, we divide them into a training part and a testing part. The testing part will be used to evaluate various machine learning algorithms.

We first evaluate the performances of three classification algorithms, Naïve Bayes (NB) [18], Support Vector Machine (SVM) [13] and Rocchio's algorithm [12]. Our experiment shows that the SVM algorithm outperforms the others. Then we examine the performances of two feature selection algorithms, Chi-square (CHI) and Information Gain (IG) [32]. Our experiment shows that IG outperforms CHI generally. We gain the best performance, about 92%, on all the performance measures, including precision, recall and F1, while 70% features are selected by using IG. However, classification performance does not change significantly after reducing a large amount of features, hence high efficiency with only a little loss of performance. The experiments suggest that among the existing machine learning algorithms, SVM classification algorithm and IG feature selection algorithm are the best choices for blog data.

Subsequently, we conduct three studies. The first study is on emotion and topic classification of blog articles. Research on emotion and mood analysis in text is becoming more popular recently [3] [17] [20]. Research on topic analysis on blog data is also very common recently [1] [5]. Those researches can enable new textual access approaches on blogs, e.g., classifying search results by emotions or topics, identifying blogger's interests, identifying communities, clustering, and so on. To classify among the affective articles, we consider two types of emotional tendencies in this paper: positive and negative. Our experiments on classifying into these two types of emotions show that we can improve by more than 16% on precision through filtering out the informative articles without decreasing the recall metric.

We then apply the above results to blog search. Generally speaking, approaches of information-access should be adapted to the type of available information. Users with different intentions may have different information needs even when they submit the same query to a search engine. In this paper, we propose an intent-driven blog-search engine. We derive the informative sense of a blog article from a confidence value that ranges from -1 (strong affective intent) to 1 (strong informative intent), and resort the search results according to the mixed scores of their informative values and original ranking values. In this way, users can view their search results according to their different intents.

Finally, we conduct a study on automatic detection of high-quality blogs. One of the reasons for people to read blogs is that it is a new source of news [16]. The well-written blogs that contain more informative articles are more attractive to readers. Intuitively, we can measure the quality of a blog by calculating the percentage of informative articles. More informative articles might mean higher quality. Detecting high-quality blogs automatically by their contents can improve the blog-search engine and can also help blog service providers attract visitors.

The rest of this paper is organized as follows. In Section 2, related work is presented. In Section 3, we give the definitions of informative articles and affective articles. In Section 4, machine learning algorithms are described in details. Section 5 presents our experiments and Section 6 gives application studies. In Section 7, we conclude our work and discuss the future work.

2. RELATED WORK

Two types of previous work are relevant to our work. One is about blog analysis, and the other is about subjectivity recognition.

2.1 Weblog Analysis

For blog analysis, there are currently two major lines of research. One line focuses on the interlinking structures of blogs. For example, Kumar et al studied how to represent the growth of blog communities and proposed a method to discover the evolution of communities [14] [15]. The other line focuses on the content of blogs. One branch is about the temporal/spatial analysis on blog contents. For example, Gruhl et al. [5] proposed a model for information propagation on blogs. Glance et al. [4] described a way to discover trends across blogs. Mei et al. [19] proposed a probabilistic approach to learn spatio-temporal theme patterns on blogs. Another branch focuses on the sentiment and emotion mining of bloggers. Durant and Smith [3] investigated existing technologies and their utility for sentiment classification on blog articles. Leshed and Kaye [17] presented a system that learns to recognize emotions based on textual resources using blog articles. Mishne [20] conducted serial experiments on mood classification of blog articles.

Our work is quite different from the existing content-based work on blogs. The existing work about sentiment and emotion classification of blog articles can be considered as a kind of domain-specific learning. [3] focused on political blogs and extracted the articles of two types of voice (Conservative vs. Liberal) on a specific topic which contains distinct sentiments. [17][20] both conducted their work on the collection of articles with bloggers' emotions. Our work makes it possible to process all the articles in the whole blog space. We want to classify blog articles into informative or affective category. Furthermore, our work can be used as a pre-processing of existing work to improve efficiency, which will be presented in the following.

2.2 Subjectivity Recognizing

Much research work on subjectivity has been done in the field of Natural Language Processing (NLP). In Weibe and Bruce' work [2] [30], a corpus of 1,001 sentences of the Wall Street Journal Treebank Corpus was manually labeled with subjectivity categorization. Riloff and Wiebe [26] presented a bootstrapping process that could learn linguistically rich patterns for subjective (opinionated) expressions. Wiebe and Wilson [31] proposed three types of potential subjective element: unique word, adjective /verb and fixed-n-gram of words. Wiebe [29] identified strong clues of subjectivity using the results of a method for clustering words according to distributional similarity, seeded by a small amount of detailed manual annotation. Turney [27] classified reviews by calculating the average semantic orientation of the phrases in a review that contain adjective or adverbs.

Our work is also different from theirs. Generally, the existing work mainly studied the linguistic features and focused on phrases or sentences. It may not be practical to apply them to our domain where millions of articles are to be processed. Moreover, the definition of "informative" could be more general than that of "objective". In this paper, we use the technique of text classification to solve the problem and will show its effectiveness and efficiency.

3. DEFINITION OF INFORMATIVE AND AFFECTIVE ARTICLES

To define informative articles and affective articles, we first do a survey among the users who usually participate in the activities in blogs, including which topics and what kind of contents they prefer to read. According to the survey, the main descriptions of informative articles and affective articles can be summarized as follows.

Informative article. The contents of this genre of articles include:

- News that is similar to the news on traditional news websites.
- Technical descriptions, e.g. programming techniques.
- Commonsense knowledge.
- Objective comments on the events in the world.

Affective article. The contents of this genre of articles include:

- Diaries about personal affairs.
- Self-feelings or self-emotions descriptions.

4. ALGORITHMS

We wish to gauge the effectiveness of known algorithms on text classification. Two types of technologies are considered to determine their applicability in our domain. One is classification algorithms. The other is feature selection algorithms.

4.1 Classification Algorithms

Naive Bayes Classifier (NB). The Naive Bayes classifier which is based on a simple application of Bayes rule is a simple but effective machine learning algorithm. It performs very well while being applied to text classification [18] [24]. Applying it to text classification, it can be presented as:

$$P(c \mid d) = \frac{P(c) \times P(d \mid c)}{P(d)}$$

where c denotes category and d denotes document. $P(c)$ is the prior distribution of a category. NB can be constructed by seeking the optimal category which maximizes the posterior $P(c|d)$.

Assume all of the attribute values are independent to the given category label. In addition $P(d)$ is a constant for every category c, we can get:

$$c^* \propto \arg\max_{c \in C} \left\{ P(c) \times \prod_{j=1}^{K} P(w_j \mid c) \right\}$$

where a document d is represented by a vector of K attributes which are treated as words appearing in the document d, $d=(w_1, w_2, \ldots, w_k)$. $P(w_j|c)$ stands for the probability that the word w_j occurs in a category c in training data, and Laplace smoothing method is usually chosen to estimate it to overcome the zero-frequency problem.

Support Vector Machine (SVM). The support vector machine (SVM) is a powerful supervised learning algorithm developed by Vapnik [28]. It has been successfully applied to text classification and performs very well [13]. In its simplest form, the goal of a linear SVM is to find the hyper-plane which can split the positive examples from the negative examples by maximizing the distance between the nearest of the positive and negative examples to the hyper-plane. Using a kernel function, the nonlinear SVM maps the input variables into a high dimensional space, and linear SVM can be applied in that space. In the binary classification, the corresponding decision function is:

$$f(\vec{x}) = sign(\sum_{i=1}^{n} a_i y_i K(\vec{x_i}, \vec{x}) - b)$$

where K is a kernel function. Polynomial kernel, Gaussian RBF kernel and sigmoid kernel are the typical kernel functions usually used in SVM. Nonlinear SVM with Gaussian RBF kernel is used in our work.

Rocchio Classifier. The Rocchio classifier relies on an adaptation to text classification of the well-known Rocchio relevance feedback algorithm in Information Retrieval [12]. It is a category profile based classifier. Rocchio's method computes the profiles of all categories in training step by means of the formula:

$$\vec{c_j} = \alpha \frac{1}{|c_j|} \sum_{\vec{d} \in c_j} \frac{\vec{d}}{\|\vec{d}\|} - \beta \frac{1}{|D - c_j|} \sum_{\vec{d} \in D - c_j} \frac{\vec{d}}{\|\vec{d}\|}$$

where $\vec{d}$ denotes document with terms weighted by the TFIDF scheme, and D denotes training set. $|c_j|$ is the number of documents in the category c_j, while $|D-c_j|$ is that not in c_j. α and β are the parameters to adjust the relative impact of positive and negative training examples. As recommended, α =16 and β =4 are used [12] in our work.

4.2 Feature Selection

Here we mainly refer to Yang's [32] descriptions for Information Gain and χ^2 statistic.

Information Gain (IG). IG is usually used to measure the effectiveness of an attribute in classifying the training data in machine learning. It measures the number of bits of information obtained for category prediction by knowing the presence or absence of term in a document. The IG of a term t can be calculated as follows:

$$\begin{aligned} G(t) = & -\sum_{i=1}^{m} P(c_i) \log P(c_i) \\ & + P(t) \sum_{i=1}^{m} P(c_i \mid t) \log P(c_i \mid t) \\ & + P(\bar{t}) \sum_{i=1}^{m} P(c_i \mid \bar{t}) \log P(c_i \mid \bar{t}) \end{aligned}$$

where $\{c_i\}_{i=1}^{m}$ denotes the set of categories and $P(c_i)$ is the distribution of a category c_i calculated by using all terms in that category in this paper. $P(c_i \mid t)$ is the distribution of the category c_i given the term t and $P(c_i \mid \bar{t})$ is the distribution of c_i calculated by using the amounts of all terms except t. This function can measure

the goodness of a term globally with respect to all categories on average.

χ^2 statistic (CHI). The lack of independence between *t* and *c* can be measured by using the χ^2 statistic. The χ^2 score of a term to a category can be calculated by using:

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}$$

where *A* is the co-occurrence time of *t* and *c*, *B* is the occurring time of *t* without *c*, *C* is the occurring time of *c* without *t*, *D* is the number of times neither *c* nor *t* occurs, and *N* is the total number of documents. The score of a term for ranking in feature space can be calculated by using:

$$\chi^2_{avg}(t) = \sum_{i=1}^{m} P(c_i)\chi^2(t,c_i)$$

In the selection phase, we rank the terms by their scores and select the top *N* terms as a feature set.

5. EXPERIMENTS

5.1 Data Set

Since there are no standard dataset on blogs, in this experiment we collected the data from the Web, which are in Chinese.

We adopt a human-evaluation approach to build the data set. We ask labelers to judge the category of a blog article according to the definitions of informative articles and affective articles. Among the 5,000 articles crawled from MSN Space [8], 3,547 are labeled as affective articles and 1,109 are labeled as informative articles, while the others are filtered because of the encoding problem of their contents. In order to balance the articles in both categories and reduce the workload of human labeling, we select 2,200 articles from Sohu.com Directory as informative articles. Sohu.com is one of the most popular websites in China [10]. The selected articles mainly are news or commonsense knowledge or objective comments about 22 different topics mapping to Sohu's Directory. They are proper to the definition of informative articles. Because our aim is to classify blog articles, those articles selected from Sohu site are only used as a part of the training data. We divide those data labeled by human into training and testing parts. The global statistic of the data set is summarized in Table 1[1].

In data-tokenization phase, we perform word segmentation and removal of the stop words.

Table 1. Statistics of Data Set

	Total Count	Training Count	Testing Count
Information	3309	2859	450
Affectiveness	3547	2920	627

5.2 Evaluation Measurer

In this paper, we employ the standard metrics to evaluate the performances of the classifiers, including precision, recall and F1-measure [25]. Precision is the percentage of correct assignment to all documents assigned to one category. Recall is the percentage of correct assignments to all the documents that should be assigned to a category. F1 is the harmonic average of precision and recall as shown below:

$$F1 = \frac{2 \times P \times R}{P + R}$$

In our experiments, macro-average gives an equal weight to every category and micro-average gives an equal weight to every document are all used to evaluate the average performance across multiple categories for F1 but only macro-average Precision and Recall are presented.

5.3 Comparing Different Classification Algorithms

Table 2 shows the performances of three classification algorithms. It can be seen clearly that SVM outperforms the others on all the measures, precision, recall, MacroF1 and MicroF1. Rocchio performs worst. Therefore, only SVM is considered in the remainder.

Table 2. Performances of three classification algorithms

	Precision	Recall	MicroF1	MacroF1
NB	0.890	0.841	0.864	0.852
SVM	0.922	0.910	0.918	0.915
Rocchio	0.860	0.727	0.772	0.730

5.4 Comparing Different Feature Selection Algorithms

Table 3 shows the performances of SVM on different amount of features for IG and CHI-square respectively. We can see that in general IG outperforms CHI on all measures, precision, recall, MicroF1 and MacroF1. In addition, we gain the best performance by using IG when 20,000 features are selected which are about 70% features. Nevertheless, the performance is not sensitive to the amount of features and a relatively high performance also can be achieved by using a little amount of features. In our view, the reason could be that the representative features of the two categories are quite different. A few of features may contain enough information to discriminate the two categories. It also suggests that there are surely two differentiated genres of contents in blogs. Table 4 shows the top 20 representative features selected by using IG of the two categories respectively.

To compare the efficiency of classification approaches with different amount of features, we record the time costs that they classify 500 articles. Every article is 3.434 KB in length on average. Table 5 shows the records for three feature sets. We can see that there is significant improvement on efficiency when using a little amount of features, e.g. 3,000, comparing to use a larger amount. It suggests that in real application scenario much higher efficiency can be achieved by reducing much more features, which brings only a little loss on performance.

The above experiments suggest that among the machine learning algorithms selected in this paper, SVM classification algorithm and IG feature selection algorithm are the best choices.

[1] This data set can be obtained from http://www.apexlab.org/apex_wiki/ia-blogdata.

Table 3. Performances on different feature set

Feature Count		Precision	Recall	MicroF1	MacroF2
100% (28579)	IG	0.922	0.910	0.918	0.915
	CHI	0.922	0.910	0.918	0.915
70% (20000)	IG	**0.932**	**0.916**	**0.926**	**0.922**
	CHI	0.926	0.914	0.922	0.919
52% (15000)	IG	0.925	0.910	0.919	0.916
	CHI	0.922	0.910	0.918	0.915
20% (6000)	IG	0.923	0.908	0.917	0.914
	CHI	0.912	0.912	0.915	0.912
10% (3000)	IG	0.914	0.903	0.911	0.907
	CHI	0.905	0.900	0.905	0.902
3% (1000)	IG	0.906	0.881	0.895	0.889
	CHI	0.901	0.895	0.901	0.897

Table 4. Top 20 representative features of each category

Informative Category	Affective Category
摄影 (Photography) 工具(Tool)	真的(Genuine) 觉得(Feel)
图文(Figure & Text) 地图(Map)	朋友(Friend) 感觉(Feeling)
报告(Report) 动漫(Cartoon)	开心(Happy) 事情(Thing)
数码(Digital Device) 专家(Expert)	喜欢(Like) 幸福(Blessing)
建设 (Construct) 文学(Literature)	一起(Together) 晚上(Evening)
美术(Painting) 资讯(Information)	今天(Today) 快乐(Joy)
奥运(Olympiad) 房产(Real Estate)	妈妈(Mother) 心情(Emotion)
经济(Economy) 资源 (Resource)	记得(Remember) 希望(Hope)
影视(Movie & Television) 艺术(Art)	明天(Tomorrow) 日子(Day)
军事(Military Affairs) 工程(Project)	哭(Cry) 每天(Everyday)

Table 5. Efficiency comparison

Feature Count	Time (ms)
10% (3000)	18111
70% (20000)	21887
Full (28579)	25324

6. STUDIES ON FURTHER APPLICATIONS

In this section, we conduct three empirical studies on the applications of above information-affectiveness classification approach.

6.1 Emotion and Topic Classification

Automatic recognition of human emotion has long been a hot research topic. In recent years, with the dramatic increase of blog data, plentiful of free textual data with people's emotions or feelings can be obtained that attract many researchers to the domain of blogger's emotion recognition [3] [17] [20]. They want to predict the most likely state of mind with which a blogger was posting an article. In general, they all address this task by classifying blog articles to some predefined emotional categories. As we know, there naturally are two genres of contents in blogs. We assume that informative articles do not express personal emotions. Therefore extracting the affective articles from the whole textual data at first could be the fundamental step of this research work. It can help to build a corpus with pure emotional articles. By this step, emotion recognition can easily be applied to the whole blog space. Content-based topic analysis on blogs is also a hot research topic in recent years [1] [5]. Extracting the informative articles may also benefit this type of research work. In this section, we will study the application of information-affectiveness classification approach on emotions and topics analyses on blogs.

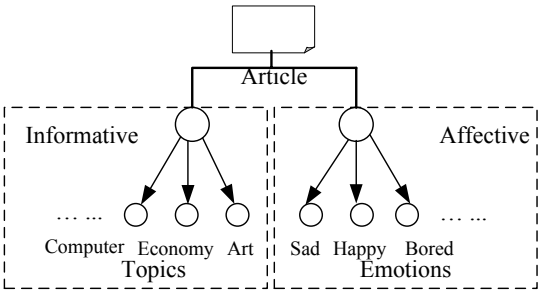

Figure 1. Two approaches for two applications

Figure 1 describes our approach that uses information-affectiveness classification as the first step of emotion and topic classification tasks. The articles which are classified to informative category could be further classified to different topics. The articles classified as affective category can be further classified to different types of emotions. To examine the effectiveness of this approach, we conduct the following experiment.

2494 blog articles are manually labeled into two emotion tendencies, positive and negative. They are used to train a binary classifier for emotion classification. Then we randomly select 75 blogs from MSN Space with 1,303 articles totally as testing data which are also labeled manually. For the selected blogs, not all articles can give emotion tendencies of the bloggers because of the existence of informative articles and the ambiguity of emotion tendency in some articles. We removed from testing data the articles describing the bloggers' feelings or emotions but with ambiguous emotion tendency. The statistics of the training and testing data used in this experiment are summarized in Table 6[2].

Table 6. Data set used for emotion classification

	Positivity	Negativity	Information
Training Data	1256	1238	
Testing Data	311	457	248

[2] This data set can be obtained from http://www.apexlab.org/apex_wiki/ia-blogdata.

To show the effectiveness of information-affectiveness classification approach for this type of emotion classification, we compare the performances of two emotion classification approaches, one firstly filtering the informative articles by using information-affectiveness classification (Ⅰ-Approach) and the other not (Ⅱ-Approach). Table 7 shows the experimental results. We first use only positive and negative articles as testing data to examine the performance of the trained binary classifier. We can see that it has a 0.788 precision and a 0.797 recall. Then we apply the classifier to all testing articles (including informative articles). We can see that filtering the informative articles firstly improves the precision of emotion classification significantly and brings a little loss on recall. The precision is almost equal to the one when testing on positive and negative articles.

Table 7. Comparison results for two emotion classification approach

	Precision	Recall
Testing only on positive and negative articles	0.788	0.797
Testing on all articles		
Ⅰ-Approach	0.762	0.785
Ⅱ-Approach	0.596	0.797

The effectiveness of information-affectiveness classification approach to the content-based topic analysis on blogs, e.g. topic classification, could be similar to above emotion classification task, so in this paper further experiments will not be conducted.

6.2 Browsing in the Weblog Space

6.2.1 Intent-driven Weblog-Search Engine

The blog space has been expanding in an incredible speed in recent years, which has become a new source of information. Many people have great excitement for participating in the activities in blogs. They may be bloggers who publish their diaries about their feelings, thoughts or the interesting events or information accessors who surf in the blog space. To access useful information easily in this space, there are several kinds of blog-search engines, e.g. [7], where users can get what they want by submitting a query. In general, after a query is submitted, existing blog-search engines usually return relevant blog articles or blogs sorted by relevance or date. Currently, blog search is at the state of Web search in 1997 [6]. As we know, there are two genres of contents in blogs. Different people may have different requirement for the same query. For example, for query "IBM", someone may prefer to find some news about IBM Company or its products or services and another may prefer to find some bloggers' comments or feelings on IBM Company or its products or services. Information-access mechanism should be adapted to the type of available information. Therefore, it is highly desirable to provide an intent-driven blog-search engine to achieve users' different requirements while they are browsing in the blog space. In this paper, we consider the two genres of blog contents as the two retrieval intents and an intent-driven blog-search engine will be given in this section[3].

Figure 2. Intent-driven blog-search engine and the top 5 search results for query "IBM" [4]

We derive the informative sense of a blog article from a confidence value that ranges from -1 (strong affective intent) to 1 (strong informative intent), and re-rank the search results according to the mixed scores of their informative values and original ranking values. Figure 2 is the screenshot of our intent-driven blog-search engine demo and the top five search results (blog articles) for query "IBM". A slide bar is provided for users to quickly access the content that they prefer by adjusting the position of the pivot according to their retrieval intents. The position of the bar corresponds to the value of parameter λ which is used to recalculate the scores of search results for re-ranking. When the bar is in the middle λ is set to be 0, and when it is in the left part and right part, λ will be set to be in range (0,1] and range [-1,0) respectively. We use Equation (1) to calculate the mixed scores.

$$S_{mixed} = \lambda \cdot S_{if} + (1 - |\lambda|) \cdot S_{origin} \quad (1)$$

where S_{if} is the confidence value for informative intent, and S_{origin} is the original value used to sort search results by relevance or by date.

In this demo, only relevance value is considered. Briefly, users who prefer to read the informative articles can drag the bar to the left part, and users who prefer to read the affective articles can

[3] The demo of the search engine can be accessed via http://infoset.apexlab.org.

[4] The results have been translated into English at APPENDIX.

drag the bar to the right part. This type of intent-driven blog-search engine generally improves user experience and satisfaction.

Figure 3. Exploring the blog named "Jolay's Home" [5]

In addition, a user could also have different exploring intents to different blogs. A user who is interested in a domain may usually prefer to read some informative articles while exploring a blog of an expert of that domain. For example, when a programmer is exploring a blog of a programming expert, he / she usually prefers to read some technical articles about programming. A user who is exploring a blog maintained by a company or an organization may prefer to read some news about the company or organization. On the other hand, a user may usually prefer to read the affective articles while exploring the blogs of his / her close friends. Users with varying requirement can also be satisfied by using our intent-driven blog-search engine. Figure 3 shows the articles of a blog named "Jolay's Home". Those articles are sorted by date by default and we have re-ranked them by dragging the bar to the left-most end to upgrade informative articles. We can see that there are two genres of articles. Users can resort the articles according to their exploring intents by adjusting the slide bar. They can access variant genres of contents quickly and easily.

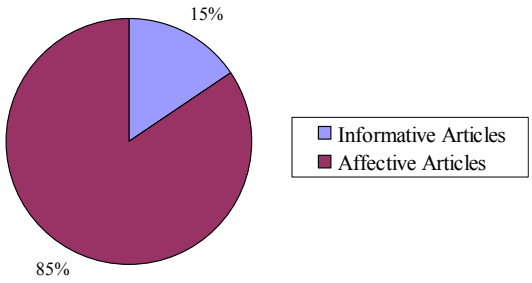

Figure 4. Distribution of informative articles and affective articles on 99,059 blog articles

6.2.2 Analysis for the Distribution of Two Genres of Articles

To measure the distribution of informative and affective articles in the blog space, we crawled 6,319 blogs with 99,059 articles from MSN Space. We apply the information-affectivness classifier trained by using all the labeled data to those articles. Figure 4 shows the distribution of the two genres of blog contents. We can see that the distribution is unbalanced. Informative articles only occupy 15%. This suggests that these blogs are mainly platforms of personal-feeling expressions. That result is consistent with the survey in [21], where it was found that people are mainly posting their personal experiences. Meanwhile the percentage of informative articles indicates that this genre of content in blogs should not be ignored. Due to the huge amount of articles in the whole blog space, the informative articles are still one prevalent informative source in the Web.

6.3 Detecting High-quality Blogs

Similar to the existence of the two genres of blog articles, there are also two genres of blogs. One genre tends to contain more informative articles. The other genre tends to contain more affective articles. Here, we will refer to them as informative blogs and affective blogs respectively. In general, informative blogs can usually attract more readers than affective blogs. On one hand, informative blogs are similar to the traditional portal websites that provide news, knowledge and different kind of information for readers. On the other hand, they are platforms of communication where readers can directly contact the bloggers who are usually some kinds of experts of variant domains. Those blogs are usually recommended by the readers to more Web users. However, affective blogs are usually the platforms of self-expression mainly for personal feelings or emotions. In this paper, we use the percentage of informative articles to all articles in a blog to measure the quality of the blog. Higher percentage means higher quality. It is highly desirable to detect high-quality blogs. Blog service providers can attract and maintain visitors by recommending high-quality blogs. Blog-search engines can also rank the blogs based on the quality of a blog in the search results.

[5] The results have been translated into English at APPENDIX.

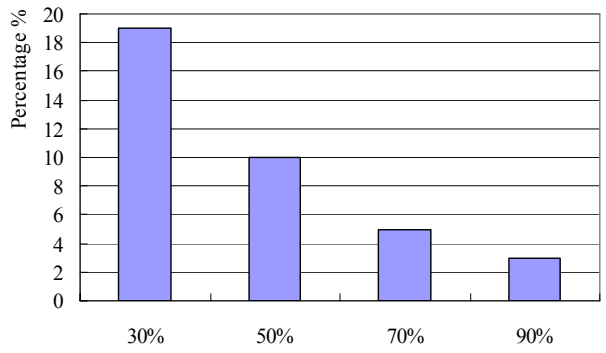

Figure 5. Distribution of blogs with different levels of quality on 6,319 blogs

To measure the distribution of blogs with different levels of quality, we apply the information-affectiveness classifier trained by using all the labeled data to the 6,319 blogs. Figure 5 shows the distribution of four-level blogs in which the percentage of informative articles is no less than 30%, 50%, 70%, 90%, respectively. The vertical axis indicates the percentage of one level of blogs versus all blogs. We can see that this ratio decreases significantly as the quality level increases. About 19 percent of blogs have no less than 30% informative articles. This suggests that the qualities of most blogs are low and those blogs are mainly platforms of personal-feeling expressions, which is again consistent with the survey in [21] where they have found that most bloggers are focused on describing their personal experiences to a relatively small audience of readers.

7. CONCLUSION AND FUTURE WORK

In this paper, we have addressed the task of separating the two genres of blog articles, informative articles and affective articles. This task is considered as a binary classification task solved by text classification techniques. By using Information Gain to select 70% features and using Support Vector Machine as the classification algorithm, we have improved on all the performance measures, including precision, recall and F1. It is also found that reducing a large amount of features does not hurt the performance but significantly improves on efficiency.

Furthermore, we have studied the applications of above information-affectiveness classification solution to other blog related work. According to our study, emotion and topic classification of blog articles can be improved by firstly using information-affectiveness classification approach. We have gained more than 16% improvement on precision when classifying blog articles to two types of emotional tendencies, positive and negativity. In addition, an intent-driven blog-search engine is proposed where the search results can be re-ranked according to whether they are informative oriented or affective oriented. The search engine technique is applicable to retrieval of articles both in the whole blog space and in a single blog domain. We have used the percentage of informative articles in a blog to measure the quality of the blog which can help to search for high-quality blogs.

We plan to extend our work to address the following issues. (1) We wish to obtain more data. Building a much large data set and reducing the labeling cost and subjectivity by using semi-supervised learning techniques are on our schedule. Existing approaches are to be applied on the data using other languages. (2) We wish to improve the classification performance even further. Combining our method with other techniques from pattern recognition is also of our interest.

8. ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for helpful comments on this work. This work is supported in part by National Foundation of Natural Science of China (No. 60473122). It is also supported in part by Google.

9. REFERENCES

[1] J. Bar-llan. An Outsider's View on "Topic-oriented" Blogging. In Proceedings of the Alt. Papers Track of the 13th International Conference on World Wide Web, papers 28-34, May, 2004

[2] R. Bruce, and J. Wiebe, Recognizing Subjectivity: A Case Study of Manual Tagging. Natural Language Engineering, 2000.

[3] K.T. Durant and M.D. Smith. Mining Sentiment Classification from Political Web Logs. In Proceedings of Workshop on Web Mining and Web Usage Analysis of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (WebKDD-2006). August, 2006.

[4] N. Glance, M. Hurst, and T. Tornkiyo. Blogpulse: Automated Trend Discovery for Weblogs. In Proceedings of WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004

[5] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information Diffusion Through Blogspace. In Proceedings of the 13th International Conference on World Wide Web, pages 491-501, 2004

[6] M. Hodder. Live Web Search. http://www2.sims.berkeley.edu/courses/is141/f05/schedule.html

[7] http://blogsearch.google.com/

[8] http://spaces.live.com/

[9] http://www.china.com.cn.

[10] http://www.sohu.com/

[11] http://www.technorati.com

[12] T. Joachims, A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization, In Proceedings of 14th International Conference on Machine Learning (ICML-97), pages 143-151, 1997.

[13] T. Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features, In Proceedings of the 10th European Conference on Machine Learning (ECML-98), pages 137-142, 1998.

[14] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the Bursty Evolution of Blogspace. In Proceedings of the 12th International Conference on World Wide Web, pages 568-576, 2003.

[15] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. Structure and Evolution of Blogspace. Commun. ACM, 47(12):35-39, 2004.

[16] J.D. Lasica, Weblogs: A New Source of Information. In We've got blog: How weblogs are changing our culture, John Rodzvilla (ed). Perseus Publishing, Cambridge, MA, 2002. Also http://www.ojr.org/ojr/lasica/p1019165278.php

[17] G. Leshed and J. Kaye. Understanding How Bloggers Feel: Recognizing Affect in Blog Posts. In Proceedings of Conference on Human Factors in Computing System 2006 extended abstracts on Human factors in computing systems, pages 1019-1024, April, 2006.

[18] A. Mccallum and K. Nigam, A Comparison of Event Models for Naïve Byaes Text Classification", In Proceedings of AAAI-98 Workshop on "Learning for Text Categorization", pages 41-48, 1998.

[19] Q. Mei, C.Liu, H.Su, and C.Zhai. A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs. In Proceedings of the 15th International Conference on World Wide Web, 2006.

[20] G. Mishne. Experiments with Mood Classification in Blog Posts. In Style 2005- 1st Workshop on Stylistic Analysis of Text for Information Access, at SIGIR 2005, 2005.

[21] Pew Internet and the American Life Project. http://www.pewinternet.org/PPF/r/186/report_display.asp

[22] Pew Internet and the American Life Project. 2005. http://www.pewinternet.org/trends/Internet_Activities_12.05.05.htm.

[23] Pew Internet and the American Life Project. 2006. http://www.pewinternet.org/trends/Internet_Activities_7.19.06.htm

[24] J.D.M. Rennie, L. Shih, J. Teevan, and D.R. Karger, Tackling the Poor Assumption of Naïve Bayes Text Classifiers, In Proceedings of the 20th International Conference on Machine Learning (ICML-2003), Washington DC, USA, 2003.

[25] C.J. van Rijsbergen. Information Retrieval. Buterworth, London, 1979, 173-176.

[26] E. Riloff, and J. Wiebe. Learning Extraction Patterns for Subjective Expressions. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03), 2003

[27] P. Turney, Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceedings of 40th Meeting of the Association for Computational Linguistics (ACL-02), 2002.

[28] V. Vapnik, Principles of Risk Minimization for Learning Theory, In D.S. Lippman, J.E. Moody, and D.S. Touretzky, editors, Advances in Neural Information Processding Systems, Morgan Kaufmann, pages 831-838, 1992.

[29] J. Wiebe. Learning Subjective Adjectives from Corpora. In Proceedings of the National Conference on Artificial Intelligence 2000 (AAAI-2000), 2000.

[30] J. Wiebe, R. Bruce, and T. O'Hara. Development and Use of a Gold Standard Data Set for Subjectivity Classification. In Proceedings of 37th Meeting of the Association for Computational Linguistics (ACL-99), 1999.

[31] J. Wiebe, and T. Wilson. Learning to Disambiguate Potentially Subjective Expressions. In Proceedings of the 6th conference on Natural language learning, pages 1-7, 2002.

[32] Y, Yang, and Pedersen, J.O, A Comparative Study on Feature Selection in Text Categorization. In: Proceedings 14th International Conference on Machine Learning (ICML 97), pages 412-420.

APPENDIX

Translation for Figure 2

	Informative Sense	Snippets
1	1.00	The catalogue of IBM certification: DB2 Database Administrator DB2 Application Developer MQSeries Engineer VisualAge For Java …
2	-0.94	Crazy Me! I have hesitated between Acer and smuggled IBM for one week. I wouldn't have taken into account the price, quality or service if I had enough money …
3	1.00	Selling IBM laptop, t22p3-900, 256M30G, dvd S3/8M, independent accelerating display card. 3550 YUAN. (Post fee not included) .Please contact 30316255. We guarantee the quality. This product is only sold within Tianjing city ...
4	-0.35	I got a laptop from my friend this week. Although outdated, it is still a classical one in IBM enthusiast's mind. There are many second hand IBM laptops in the market. Although I have sold many IBM laptops …
5	-0.53	Doctor said that I should make more preparations mentally. You have stayed with me for three years, leaving without any words. Do you feel fair for me? Do you remember the moments we were together? You are heartless, I hate you! ...

Translation for Figure 3

	Informative Sense	Snippets
1	1.00	The catalogue of IBM certification: DB2 Database Administrator DB2 Application Developer MQSeries Engineer VisualAge For Java …
2	1.00	Biao Lin is a military talent. Stalin called him "thegifted general". Americans called him "the unbeaten general". Chiang Kai-shek called him "devil of war". Biao Lin is a special person in modern history …
3	0.99	Microsoft's hotmail can only be registered with suffix "@hotmail.com" by default. You can register @msn.com by visiting…
4	0.95	Yi Shang is still sending the file to me. I will practice it later. 1. Start up Instance (db2inst1) db2start; 2. Stop Instance (db2inst1) db2stop …
5	0.84	Name: Lei Zhang. Student number: 5030309959. Class number: 007. The analysis and review about the tendency of Jilin Chemical Industry' stock in 2005. Date, Increasing and Decreasing ranges, Open Price, Close Price, Amount of deals …
6	0.01	Recently I like reading the Buddhist Scripture. I can learn philosophies in it. It makes me comfortable. It is from ...
7	-0.11	It's out of my mind when I first saw it. The water seemed to be exuding from the building. There was much water on the floor of education building. Water was all around us, anywhere you can touch had water. …
8	-0.51	I read an article about the last emperor Po-yee today. I have watched "The Last Emperor" before, which realistically described his life without losing artistry. His love impressed me. As an emperor, he can't choose the one he loved …
9	-0.53	She is 164-166 cm in height with white skin, black hair and long limp leg. I like the girl who has long hair and likes sport and dancing. I like sweet girls. …
10	-0.94	I have many things to do at the end of this semester. There are five final examinations, Discrete Mathematics, Communication Theory, Architecture of Computer, Algorithm and Law. I know little about them. OMG! Only four weeks are left. There are also two projects, Compiler and Operation System. Complier can be easily completed but Operation System …