Journal of Computer Science 8 (2): 251-258, 2012
ISSN 1549-3636

An Improved Approach for Topic Ontology Based Categorization of Blogs Using Support Vector Machine

[1]Subramaniyaswamy, V. and [2]S. Chenthur Pandian
[1]Department of Computer Science and Engineering,
Selvam College of Technology, Namakkal, Tamilnadu, India
[2]Mahalingam College of Engineering and Technology, Pollachi, Tamilnadu, India

Abstract: Problem statement: Information search, collection and categorization from the blogosphere are still one of the important issues to be resolved. Mainly, the blogs assist the variety of interesting and useful information. Because of its increasing growth, blogs can not be categorized effectively. Therefore it is difficult to find relevant topics from the blogs. Hence blogs need to be categorized topically to make easy for readers. **Approach:** Blog contents are associated with a set of predefined topic ontology keywords. This study proposes categorization of blogs to facilitate easy identification of user expected topic from the massive collection of blogs. Tags, page contents were collected as inputs from the blogs and the blogs were categorized using Support Vector Machine (SVM) algorithm. Most frequent occurrences of topic ontological keywords are used to train the classifier. This approach has effectively improved blog categorization process using SVM. **Results:** The performance was evaluated for precision and recall for blog categorization based on topic ontology using SVM with Naive Bayes algorithm. It was proved that topic ontology assisted SVM improves the classification accuracy than Naïve Bayes algorithm. **Conclusion:** This study has effectively improved the classification of blogs based on topic ontology assisted SVM. Experiments showed the effectiveness of the blog categorization.

Key words: Blogosphere, blog content and tagging, topic ontology, machine learning, blog categorization, Support Vector Machine (SVM), naive bayes algorithm

INTRODUCTION

Blogs: Blogs emergences in late 1990s may consist of more interesting themes, personals, general knowledge, products, political annotations, communication, entertainment related topics. Blogs are maintained and authored by individuals. Blogs are used for knowledge sharing (Frank *et al*., 2009) and it is not stable since it catches the continuing expressions. Blog is a special type of web application. People can share their interested topics by publishing through blog. Blogs have been created by many corporate companies to introduce their products and services. Companies create customers network through blogs. Blog helps people to publish their views quickly. Based on topic relevance, blogs may have links to other blogs. Nowadays above ninety percentages of interactive blogs are available on the web. In an interactive blog, readers can leave the comments and blog merges the text and images. Blogs are composed of articles and it contains different topics. Blogs are updated frequently (Bayoudh *et al*., 2008). However, it is essential to classify the blogs based on topics. Multiple entries by user are named as blog posts or posts.

This study uses topic ontology to train SVM algorithm for classifying blogs article. Blogs may contain huge variety in the quality, content and goal of blogs. Thousands of daily readers can be handled by the blog. Formal definitions frame blogs are a type of blogs that are separated and segmented via composition and type of the page content. Composition discrepancies are examined as divergence from a normative prototype and the differences in page content are seemed to be representative of sub-types. Page content is classified based on terms and its frequency (Aizawa, 2003).

Blogosphere: A division of the web is called as blogosphere and web has weblogs. Blogosphere contains dynamic and dated content. Blogosphere necessitates specific engines to search and analyze when search engines noticing and indexing the blogs. Several resources that are significant for blog search

Corresponding Author: Subramaniyaswamy, V., Department of Computer Science and Engineering, Selvam College of Technology, Namakkal, Tamilnadu, India

engines are: identifying blog sites, structure, component part and extort relevant metadata. It is important to notice and remove the spam blogs (Kolari *et al*., 2006). The blogosphere is smooth to comment spam, trackback spam, ping spam and splogs. These are engaged by spammers and blogosphere contain self hosted blogs. Slog finding and removing is a most difficult task in blogosphere.

Tags: Set of keywords or labels are called as tags that are very useful for grouping articles into large categories. Tags are attached to blog entries to assist the blog and tagging is a process of labeling. Tags assist the visitor or blogger to comment their information and allows to share with others. Tags are very useful to improve the efficacy of search engine since content categorization is done using a well-known, accessible and common vocabulary. Tags are spontaneous and easy to use (Brooks and Montanez, 2006).

The users can tag others information and tags are mainly used to represent the properties of information. Therefore, tags are important in arranging information by subject, date, month and year. Tagging provides notes, description, distribution and categorizing to the blog posts (Frank *et al*., 2009; Brooks and Montanez, 2006). Tags can be used by anybody with any sort of blog post and different topics can be interconnected by the tags. After adding the tags for the blog entry, tags can be edited, published or saved as a draft by the blogger. Many tags can be used by separating with commas to be more specific. Publishing content is an important part for bloggers. Some terms or categories are allocated to contents by the publisher and publishes it. These published contents reach all the blog readers and comments can also added by the readers (Tramullas and Garrido, 2006). Therefore the content should be taken from the referred blog page. Really Simple Syndication (RSS) is used for creating web feeds to publish frequent updated blogs. RSS presents content that an individual desires to read. Blogs are public relationship tools (Kent, 2008).

Blog categorization and their characteristics: The following tasks are associated with the blog categorization. They are.

Spam filtering: Spam blogs are filtered out by categorizing them using SVM. Spam filtering is like cops and muggers game (Drucker *et al*., 1999).

Text filtering: Categorizing the flow of received documents transmission in an asynchronous way by an information creator to an information customer (Sebastiani, 2002).

Topic routing: Topics are to be assigned to one or more categories, relevant or non-relevant.

Categorization: Blog categorization may contain the hierarchical structure of the category set (Schutze *et al*., 1995). Atom blog is created and offered based on semantic classification and comments of content published on blog (Patel and Khuba, 2009). SVM is very scalable to large set of documents in blog mining applications. Topic ontological keywords frequencies are captured to compute the probability values between each and every topic and category. Topical keywords are used to train SVM. SVM deals with multiple categories case. SVM is fairly good in blog categorization. In this process, multiple blogs are categorized and organized into multiple predefined categories. Hence, SVM is very suitable for classifying the blogs in order to organize in a meaningful structure.

Problem definition: Variety of topics can be posted by the bloggers every day. Blogs do not have any particular structure and it does not follow any formals or professionals. Information is organized randomly in a blog. Because of its increasing growth, blog cannot be categorized effectively. Therefore it is difficult to find relevant topics and categories from the blogs. Hence blogs need to be categorized topically to make easy for readers. This study enhances the categorization of blogs using SVM to provide a meaningful structure based on topic ontology. Topic ontology assists a meaningful structure and allows readers to surf or look through a blogs without confusion. Categorized blogs can be added to the categories based on the contents.

Key contributions of this study are to:

- Employ the topic ontology to organize topics onto given categories. Extract keywords from topic ontology and most frequent topic ontological keywords or topics from are taken as training set. It trains the SVM classifier efficiently.
- Collect tags and page content from the blogs and categorize the blog based on collected tags and contents using the SVM classifier.
- Finally, present the effectiveness of the proposed approach by comparing the blog categorization using topic ontology based SVM versus blog categorization using Naive Bayes algorithm

The rest of the work is organized as follows: Describe the related work in Chapter II. Chapter III presents the representation of blog categorization using topic ontology based SVM by gathering tags and page contents. Chapter IV gives an empirical evaluation to show the efficiency of the proposed blog classification method. Chapter V presents the conclusion and future study.

Related works: Blog classification has been studied extensively in literature. A gender classification of blog authors was proposed by (Mukherjee and Liu, 2010). In this gender classification, proof of concept system was proposed to classify the gender based on the blog entries. Naive Bayes classification algorithm was utilized for recognizing genders of blogs (Yan and Yan, 2006). Sentiment analysis is one of the popular blog categorization mechanisms. For classifying the sentiment, SVM is robust and it manages the noisy data (Tharp, 1973). Sentiment classification is very useful in business intelligence applications and recommender systems. It is also used to summarize the user input and feedback rapidly. Positive and negative sentiment classifications have been done by (Pang *et al*., 2002). Twitter is one of the micro blogging services on web. Sentiment messages of this micro blogging are categorized using machine learning algorithm SVM. According to the users' query sentiments are categorized as positive or negative. This type of sentiment classification is more helpful for companies and customers to observe the sentiments of products and brands. Distant supervision method is used for sentiment categorization and it contains tweets with emoticons. Naïve Bayes, Maximum Entropy and SVM provide above 80% accuracy (Go *et al*., 2009). The SVM and Multinomial Naive Bayes (MNB) classifiers are used for classifying sentiments in microblogs (Bermingham and Smeaton, 2010).

The term Folksonomy is a manual categorization method (Ohkura *et al*., 2006; Zhang *et al*., 2006). It is a sharing of tags which aimed at getting the concept that the right handling of a tag is described by the working community as opposite to being verdict by a committee. It generates the precise meaning and the blog content should be decentralized (Brooks and Montanez, 2006). The use of tags generates a Folksonomy. Using Folksonomy, users' interests can be identified and tags lists are considered. Based on time tag lists can be increased by noticing new interests and include new tags to classify and explain it. Folksonomy is very useful method for browsing weblog articles. It is low cost and requires less man power (Ohkura *et al*., 2006).

Spam is an electronic junk mail that is voluntary, unwanted, unrelated or unsuitable. SVM provides (Drucker *et al*., 1999) speed, accuracy and less time for training the data sets in categorizing e-mail as spam or non spam. A study addressed political blog classification (Jiang and Argamon, 2008) based on its political leaning. To construct political leaning classifiers, subjective sentences include two strong subjective indications based on the common Inquirer dictionary are identified. From identified subjective sentences, opinion expressions and bag of word features are extracted. Cool blog identification (Sriphaew *et al*., 2008) was proposed using topic based models. Blogs with good and valued contents are called as cool blogs. To identify the cool blogs from the vast collection of blogs on the net, the following methods are considered. Cool blogs lean to have exact topics, adequate blog entries and topic reliability. Topic model is employed to extract topic likelihood, blog entries are utilized and distance functions over topic likelihood is used to estimate the topic reliability among blog entries. A feature amalgamation method is attains greatest effectiveness.

Blog entry selection algorithm is presented to explain the optimization task. An extrinsic evaluation method is tailored from document summarization to estimate the entry selection algorithms. Based on the accuracy of blog classification, blog entry selection algorithms are evaluated classification accuracy is attained by getting most representative entries for each blog. In anomaly, blog consists of noisy entries that are irrelevant to its main topics. Noisy entries may degrade the performance of blog mining techniques. Noisy entries are avoided in blog entry selection method. Most representative entries are chosen to the focal topic of the blog to get useful entries in representativeness mechanism. Diverse entries are selected and overall information of the entries is maximized. Repeated entries are avoided. Citation KNN algorithm is used to classify the blogs. To learn SVM classifier, popular LIBSVM is utilized in empirical evaluation. The proposed methods are efficient and promising. Blog entry selection technique reduces the computational cost effectively in order to improve the classifying task more efficiently (Zhuang *et al*., 2008). Assigning keywords to blog posts are called as tagging. Tagging is used for explaining information for personal use, placing information into largely defined categories and commenting specific articles (Frank *et al*., 2009). For grouping articles into topical categories, tagging is more efficient. Tagging is an efficient to select advertisements to display. Blogger should choose how specific the tags should be provided. Labeling and tagging are used to perform the classification. Tagging provides many methods to classify the blog. Spamming, canonicalization and ambiguity are the major problems in tagging. Support Vector Machines are generally used for all classification tasks. Classification can be topic classification, sentiment (Aizawa, 2003), email spam classifications (Drucker *et al*., 1999). SVM for categorization tasks are more effective and competence.

MATERIALS AND METHODS

Proposed approach: This chapter is structured as following: Constructed topic ontology with keyword extraction, tags, content collection and categorization

algorithm SVM. SVM classifies the blogs based on their contents and tags.

To improve the performance of blog classification, feature selection with SVM is used by finding which keywords are significant to arrange the precise category of document. The classification task is to allocate document to one or more categories based on its contents. First SVM classifier is trained using the keywords of the topics. SVM separates the blog according to the given categories.

Topic ontology: Topic ontology is a collection of topics (Zhoua *et al*., 2006). Topic ontology construction process includes assigning documents to the topics. Each document contains relevant keywords. In this topic ontology, topics are interrelated by semantic relationships. While constructing topic ontology, topics are represented as nodes and topic relations are represented as edges. Topic ontology provides means to recognize related documents for each topic. To describe topics with keywords, documents in a topic and sub topics are mentioned with positive and negative label. Each positive, negative document is a collection of keywords. Keyword frequency is calculated as number of occurrences of keyword in a document. Group of keyword frequencies are referred to as pattern. Pattern is uniquely identified by the group of keywords.

Topic ontology contains these patterns to formalize new patterns. User can edit the topics and transform which documents are assigned to topic. The user can manage all the relations among topics by adding, deleting, ordering and identifying the relations. Bag of words are employed to denote the documents in a topic which runs on the weight of the keywords. Using kfitf weighting method, keyword weights are computed. Keywords are extracted from documents of topic ontology. Each keyword may be a single word, a small phrase, or a synonym. Primarily, keywords extraction is done by summing up of all the vectors of document in the topic. Keywords with a highest weight in a topic are chosen as training set (Fortuna *et al*., 2006).

Tags and content collection: Blog post contains the information about individuals and unnecessary features are then eliminated from the blog. Online blogs contain tags and these tags are also called as keywords (Brooks and Montanez, 2006).

Content is enough for precisely categorizing the blogs. Contents are the collection of words which is called as bag of words. Each blog is represented by the vectors and each blog entry is corresponding to a single word. Vector size is the identical for the entire blogs and the blog entries in the vector are kf-itf values (Maguitman *et al*., 2010). The blog content may be personal entries or it may be useful information. Maintainer of blog is called as blogger. This study focuses on categorizing blogs to allocate different topics into categories based on the content. Blog search engine is used to collect these tags and contents from the blog. Top 11 tags and contents are collected from blogs to examine the efficacy of tags for categorizing the blog. Similarity of articles that share a tag is also calculated.

Stop words: Blogs are pre processed to filter out the pointless words from the blogs (Joachims, 1998). Stop words are also called as noise words (i.e., articles, conjunctions) and these words are not important in classifying the blogs. These words are removed from the blogs.

Stemming: To reduce the amount of different words, changes in lowercases word stemming is done in the blog categorization followed by transformation occurs (Joachims, 1998). Porter (1980) Stemmer algorithm is a powerful algorithm for stemming. Relevant topical keyword is mapped into topical categories.

Topical keywords frequency: Topical keywords frequency is calculated according to the weight of the topical keyword in the blog (Drucker *et al*., 1999). KF-ITF keyword weighting method (Aizawa, 2003) is presented to calculate the keyword frequency:

$$Kf\,\frac{Bi}{\sum_{t=1}^{n} B_t}$$

Transforming topical keyword into vector space: It is necessary to change each topical keyword into a vector since SVM is used. List of keywords present in a blog are considered. A blog is denoted as a vector and each feature corresponds to a separate topical keyword. Blogs classification is automatically assigning topics into concern categories (Tharp, 1973) and also assigning blogs one or more predefined topical categories is called as blog classification. m coordinate of the n is denoted as:

$$Kfitf(m,n) = Kf(m,n).itf(m)$$

$$Itf(m) = \log\frac{T}{tf(m)}$$

Kf (m, n) represents that occurrence of m in the n topic, T is the number of topics and tf (m) calculates the topics consisting of m at least once. Transformation of topics develops the term-topic matrix. Various lengths of topics are the same length in vector space which can be attained using topic normalization (Joachims, 1998):

$$\text{Kfitf}(m,n)\frac{\text{Kfitf}'(m,n)}{\beta}$$
$$\sqrt{\sum}\ \text{mKfitf}(m,n)$$

KF-ITF: Keyword frequency - inverse topic frequency is described as:

$$\text{Itf} = \log (m/ki)$$

Kfitf = Kf*itf is defined and weight in kfitf is accomplished by high frequency of keywords and a low frequency of the keyword. Thus, common keywords are filtered out from the blogs.

Categorization using SVM: Support vector machines are useful for blog classification since thin data and huge dimensionality are dealt with SVM. SVM is to learn and simplify an input output mapping. Inputs are the tags and contents and output is their corresponding topical categories. SVM is used to learn that how to categorize the blogs after the completion of preprocessing and transformations. SVMs have been proven as one of the most powerful learning algorithms for blog classification (Joachims, 1998). SVM classifier is used in bioinformatics gene expression and in binary classification (Bayoudh *et al*., 2008). Joachims proposes the topic classification using SVM. In SVM there are two classes involved for categorization task. Two classes are denoted using (p1, q1), ..., (pn, qn). pi denotes the vector feature and qi is linearly separable. SVM finds the weight of the vector V (Kolari *et al*., 2006):

$$\| V \|^2 \text{ is min}$$

Inputs: Tags and contents (t, c) => pi € R(t,c) ,b is a constant:

V. p+ b = 0

A hyper plane (V, b) that separates the blog, this gives the function:

f (p) = sign (V. p +b), categorizes the blog
pi*V+b> = ±1 when qi = ±1, canonical hyper plane
qi (pi*V+b)> =1, > = functional distances

Normalization of level of vector V for obtaining geometric distance from the hyper plane:

$$t, c((V,b),pi) = \frac{qi(pi * V = b)}{\|V\|} \geq \frac{1}{\|V\|}$$

Linear combination of vector V is, Li (qi (V. pi +b) -1) = 0 ($\forall_i$) c reduces the margin error. SVM performs on high dimensionality of inputs and it classifies the blogs accurately.

Topical categories: Relevant and non relevant topics are assigned to their corresponding categories. Tags specify the main topic and subtopics of the blog article and convey it by using the exact words. Blogs are classified by the topic of the blog (Maguitman *et al*., 2010). Topic ontology construction process assigns contents to the topics. Blog is classified by the topic into a predefined set of four categories. System suggests sub topics for user chosen topics and user handles the system suggested topics. User chooses the subtopics and then added automatically to ontology with relation of chosen topics. By adding, eliminating, directing and identifying the relations, user can manage all the relations between topics.

Contents are automatically assigned to a topic when it is added to the ontology via determining the keyword as topics by system. Similarity of contents is calculated using cosine similarity from the quantity and the centric of the topic. This can helps the user when looking for contents related to topics. Similarities among chosen topics and all the other topics are calculated by the system. Topics can be edited by the user in a blog. Blog consists of more topics. Topics name, relationship are mentioned and sub topics are also added to the topic ontology (Fortuna *et al*., 2006). Topic ontological blogs are means of classifying blogs based on the tags and page content (Maguitman *et al*., 2010). The topics should be defined anywhere on the blog and topics may contain several different types of tags.

Feature selection: Feature selection is a main module of machine learning applications. Here, feature selection module is used in a blog categorization to decrease the load on computational resources and to assist in stimulating the performance by removing noisy features. High frequency keywords of the topics are selected from the topic ontology.

Experiments: This chapter provides the descriptions of the inputs used for performing experiments. The machine learning algorithm SVM is trained using topical keywords and used for categorizing the blogs. Proposed method is tested empirically in order to show the effectiveness in improving classification accuracy. Topic ontology is applied in order to add the categorized blogs into given topical categories. Demonstrate that SVM method allows categorizing the blogs precisely based on the tags and contents.

Table 1: Word, tags for classifying blogs

Tags	Bag of words	Blog frequency	Topical categories	Classification accuracy using SVM (%)	Classification accuracy using NB (%)
Cricket	Bat, ball	671	Sports	94.58	75.12
Personal	Stories	778	Personal	96.73	65.18
Humor	Jokes	503	Entertainments	93.42	35.02
Marketing	Products, services	321	Business	90.12	50.02
Video games	Car race, bounce ball	211	Entertainments	93.42	35.02
Thoughts	Ideas, own stories	406	Personal	96.73	65.18
Sales	Products, services	433	Business	90.12	50.02
Sketching	Sketch, water mark	376	Arts	91.23	48.74
Shuttle cock	Cock, bat	684	Sports	94.58	75.12
Workshops	Subjects, technical	560	General	92.03	20.00
Schools, colleges	Books, students	600	Education	90.00	75.01

Experimental setup: For this experiment, training set and test data are used. Using SVM algorithm, parameters should be trained. Initially, keywords and topics from topic ontology are used as a training set. Most frequent occurrences of keywords or topics from topic ontology are used as a training set. Training set includes a sample of 100 blogs for classification. Blog contents and tags are used as a test data. Blog search engine collects the tags and contents of the blog. A single blog can have multiple tags and contents. The first feature space is blog content which is also called as bag of words. Here, vector is used to indicate each blog and blog entry is relevant to a single word in a blog. Blog entries in the vector are kf-itf values. Classifier is trained separately and tested. All blogs are structured in a topical categories arranged in alphabetical order. For each blog, blog engine provides tags, contents, blogger and categories allocated to the blog. Tags and content of the blog page are used as test data and blogs are categorized using SVM. The 11 popular tags, word of tags, blog frequency, topical categories and classification accuracy have been listed in Table 1. In this categorization process, topics are assigned to a particular category.

Bolded tags belong to topical category Sports, underlined tags are belonging to Business category. Tags and contents are effective for classification of blog. Kf-ITF method provides similar classification accuracy.

RESULTS

Experimental evaluation: Precision and recall used for calculating the classification accuracy via exact comparison of classifying blogs into predefined categories versus simple blog classification. Precision is a measure of the efficacy of retrieved tags, contents and recall is a measure of the wholeness of the retrieval procedure. To evaluate whether tags and contents direct to more precise classification results, precision and recall are calculated. After stop word removal and stemming, content has 152.8 words on average for each blog.

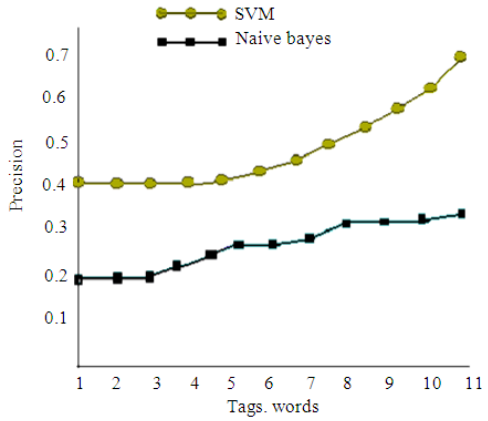

Fig. 1: Precision

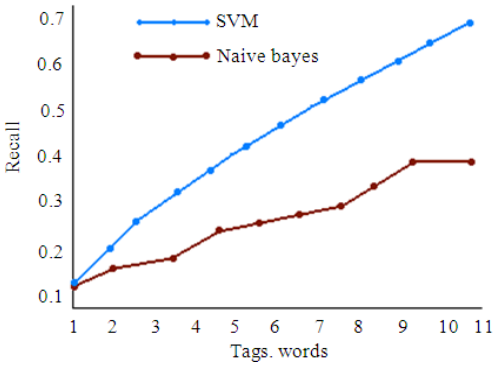

Fig. 2: Recall

The number of words in content is about double times of the 305.6 words contained in tags on average. When tags are increasing the classification accuracy is also increased when compare to a simple blog classification.

Figure 1 shows that tag together with contents achieved the best precision in classifying blog compared with NB blog classification. Precision increased along with the tag and contents of the blog. More tags, contents directed to improved precision. Figure 2 depicts recall for blog categorization using SVM and Naive Bayes algorithms.

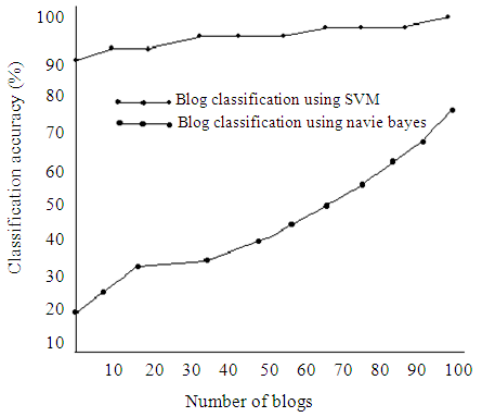

Fig. 3: Classification accuracy

DISCUSSION

Classification accuracy: The accuracy performance measure is used for evaluation. Table 1 shows the experimental results that are measured by averaging the accuracy values over all the tags. SVM provides classification accuracies of 94.58% for sports category, 90.12% of accuracy for business category. For other categories, SVM provided 96.73 and 93.42 and 91.23% of accuracies over the tags and contents based classification. It shows that the topics in topic ontology of blogs are profits higher accuracies for all topical categories. Figure 3 demonstrates increased percentage for blog classification when compared to a simple blog classification.

CONCLUSION

This study introduced the task of categorizing blogs using SVM and compared with blog classification using Naïve Bayes method.

Blog categorization process is experimented and evaluated. The experiments showed the advantages of topic ontology to assist in categorizing blogs and this improves the results of the categorization task. Supervised learning algorithm SVM performed better in categorizing the blogs. It provides better accuracy and it is well suited for blog classification. However, topic ontology is very useful to provide a semantic structure for topics in a blog. By getting tags and contents categorization of blogs into pre-defined topical categories has been done successfully. From the experimental results, tags and contents were more effective for accurate classification than the extracted features from the blog. Experimental results showed that more tags and contents lead to better classification accuracy.

REFERENCES

Aizawa, A., 2003. An information-theoretic perspective of tf-idf measures. Inform. Proc. Manage., 39: 45-65. DOI: 10.1016/S0306-4573(02)00021-3

Bayoudh, I., N. Bechet and M. Roche, 2008. Blog classification: Adding linguistic knowledge to improve the K-NN algorithm. IFIP Advan. Inform. Commun. Technol., 288: 68-77. DOI: 10.1007/978-0-387-87685-6_10

Bermingham, A. and A.F. Smeaton, 2010. Classifying sentiment in microblogs: Is brevity an advantage? Proceeding of the 19th ACM International Conference on Information and Knowledge Management, (IKM' 10), ACM, New York, USA., pp: 1833-1836. DOI: 10.1145/1871437.1871741

Brooks, C.H. and N. Montanez, 2006. Improved annotation of the blogosphere via autotagging and hierarchical clustering. Proceedings of the 15th International Conference on World Wide Web, (WWW'06), ACM New York, USA., pp: 625-632. DOI: 10.1145/1135777.1135869

Tharp, A.L., 1973. Using verbs to automatically determine text descriptors. Inform. Storage Retri., 9: 243-248. DOI: 10.1016/0020-0271(73)90092-2

Drucker, H., D. Wu and V.N. Vapnik, 1999. Support vector machines for spam categorization. IEEE Trans. Neural Networks, 10: 1048-1054. DOI: 10.1109/72.788645

Fortuna, B., D. Mladenic and M. Grobelnik, 2006. Semi-automatic construction of topic ontologies. Semantics Web Mining, 4289: 121-131. DOI: 10.1007/11908678_8

Frank, J., R. Motschnig and M. Homola, 2009. Towards an "intelligent" tagging tool for blogs. Best Practices Knowledge Soc., 49: 129-136. DOI: 10.1007/978-3-642-04757-2_14

Go, A., R. Bhayani and L. Huang, 2009. Twitter sentiment classification using distant supervision. Stanford University.

Jiang, M. and S. Argamon, 2008. Political leaning categorization by exploring subjectivities in political blogs. Illinois Institute of Technology.

Kent, M.L., 2008. Critical analysis of blogging in public relations. Public Relations Rev., 34: 32-40. DOI: 10.1016/j.pubrev.2007.12.001

Kolari, P., T. Finin and A. Joshi, 2006. SVMs for the blogosphere: Blog identification and splog detection. University of Maryland.

Maguitman, A.G., R.L. Cecchini and C.M. Lorenzetti and F. Menczer, 2010. Using topic ontologies and semantic similarity data to evaluate topical search. Universidad National del Sur.

Porter, M.F., 1980. An algorithm for suffix stripping. Program: Electronic Library Inform. Sys., 14: 130-137. DOI: 10.1108/eb046814

Mukherjee, A. and B. Liu, 2010. Improving gender classification of blog authors. Proceedings of Conference on Empirical Methods in Natural Language Processing, (EMNLP' 10), Association for Computational Linguistics, Stroudsburg, PA, USA., pp: 207-217.

Ohkura, T., Y. Kiyota and H. Nakagawa, 2006. Browsing system for weblog articles based on automated folksonomy. Tokyo University.

Pang, B. and L. Lee and S. Vaithyanathan, 2002. Thumbs up?: Sentiment classification using machine learning techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, (EMNLP' 02), Association for Computational Linguistics, Stroudsburg, PA, USA., pp: 79-86. DOI:10.3115/1118693.1118704

Patel, D.R. and S.A. Khuba, 2009. Realization of semantic atom blog. J. Comput., 1: 34-38.

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. Mach. Learn., 1398: 137-142. DOI: 10.1007/BFb0026683

Schutze, H., D.A. Hull and J.O. Pedersen, 1995. A comparison of classifiers and document representations for the routing problem. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (RDIR' 95), ACM. New York, USA., pp: 229-237. DOI: 10.1145/215206.215365

Sebastiani, F., 2002. Machine learning in automated text categorization. ACM Comput. Sur., 34: 1-47. DOI: 10.1145/505282.505283

Sriphaew, K., H. Takamura and M. Okumura, 2008. Cool blog identi? Cation using topic-based models. Proceeding of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Dec. 9-12, IEEE Xplore Press, Sydney, NSW, pp: 402-406. DOI: 10.1109/WIIAT.2008.401

Tramullas, J. and P. Garrido, 2006. Weblogs content classification tools: Performance evaluation. proceeding of the Internacional Conference on Multidisciplinary Information Sciences and Technologies, Oct. 25-28, Mérida, Spain, pp: 532-536.

Yan, X. and L. Yan, 2006. Gender classification of weblog authors. Stanford University.

Zhoua, X., Y. Lia, Y. Xua and R. Laub, 2006. Relevence Assessment of Topic Ontology. In: Advances in Intelligent IT. Y. Li, M. Looi and N. Zhong, (Eds.). IOS Press, Amsterdam, ISBN: 1586036157, pp: 44-51.

Zhuang, J., S.C.H. Hoi and A. Sun, 2008. On profiling blogs with representative entries. Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data, (ANUTD'08), ACM, New York, NY, USA., pp: 55-62. DOI: 10.1145/1390749.1390759