Semantic Tagging and Classification of Blogs

A. K. Singh[1], R.C. Joshi[2]
[1,2]Department of Electronics and Computer Engineering, Indian Institute of Technology Roorkee-247 667, Uttaranchal, India, [1] anilkdec | [2]joshifcc @iitr.ernet.in

1. Introduction

User generated content is one of the main exciting portion of the web. Blogs constitutes an important part of this portion, which is growing rapidly. It is fast becoming a tool for information sharing and social networking. A blog is a website maintained by an individual that contains commentary, descriptions of events, graphics or video. Entries in blogs are always displayed in reverse chronological order. In the recent years the uses of blogs have drastically increased. Because of its diversified use and its rapid increase, it has widely attracted researcher's attention. Blog classification is required for the relevant blog searches, grouping blogs at a higher conceptual level, and its use for blog post recommendations.

There are many classification schemes which classify the documents using supervised learning approaches and semi supervised approaches. Most of these methods don't consider the semantics of the text and rely on the text based classification methods. This work presents an approach which takes the best of "bag of words" approach and then improves it with the knowledge represented in the respective domain ontology. It uses domain ontology together with the features extracted, and subsequently refined, from the blog posts to classify the blogs. The ontology we have used in this paper is simple. It's just comprises of classes, properties, instances, subclass - super class relationship and the relationships that connects two entities. The approach that we have used not only considers the occurrences of words but also establishes the semantic relationships between them. A semantic graph is obtained from the semantically refined term vector of a blog post. The terms in the term vector are either represented as entity, or instances in the semantic graph. After a document is converted into a semantic graph of entities, the ontological classification of the entities in the graph is then analyzed in order to determine the overall categorization of the semantic graph, and as a result, of the document. The analysis of this semantic graph not only classifies the blog to some generic categories but it establishes the hierarchy, which accurately tells us to which categories the blog belongs to.

2. Related work

Because of the increasing number of blogs and their unique characteristic, blogs have received much attention from researchers and various studies have been conducted. Ni et al. [1] presented a machine learning method for classifying informative and affective articles among blogs. Chi et al. [2] proposed a novel technique that captured the structure and temporal dynamics of blog communities. Chen et al. [3] developed blog-specific search and mining models with the aim of better strengthening business intelligence in complex enterprise applications. Nevertheless, very few studies on blog categorization have been reported. In a recent study, Hu et al. [4] proposed a solution to extract representative sentences from a blog post using information hidden in its comments. Blog comments are regarded by most bloggers as vital to the interactive nature of blogs [5]. Taking into account existing blog comments could help produce more reader-oriented summaries [6]. Despite this, quantitative studies of blogs focus on post data, leaving out the comments. Zhou et al. [7] assumed that the blog entries were summaries, with personal opinion added, of online news articles that they were linked with. A summary is generated by deleting sentences that are not related to the linked news articles, ignoring the comments.

a. Problem definition

The domain of blog classification has a naturally evolved solution to it. The blogger uses representational keywords for labeling. These keywords are known as tags. These tags are provided by the blogger based on his personal understanding of the domain, and thus could be biased. Existing social bookmarking and tagging system like delicious.com, flickr.com is using tags to label a blog post. They even offer tag searching service. There are at least three barriers to utilize blog tags in classification or navigation. A substantial amount of the blogs may not be tagged, there are many orthographic or synonymous tag variations, and not all tags are informative. However the semantics of tags, and as a result the semantics of the resources, are not known and are not explicitly stated. This often hampers the resource retrieval within the individual system as well as the integration of resources in cross platform applications.

978-1-4244-9034-/10/$26.00©2010 IEEE

The machine learning approaches for blog classification uses supervised or semi supervised learning methods that involves substantial amount of human intervention in the classification task. This work focuses on the classification of blogs using domain ontology and semantically enhanced tag set of the blog post, to facilitate more meaningful blog organization and subsequently blog search.

Additionally, a comparison is done between two methods of classification, one that is based on tf idf scores computed by using blog features, and other that uses domain ontology to extract central entities that represents the main concepts of the blog. This information is used to enrich the tag set of the blog. It is then utilized to add effectiveness to search mechanism.

We have obtained domain Ontologies from the DAML library to enhance tag enrichment and categorizing a blog post. More specifically, ontology is a data model that represents a set of concepts (entities) within a given domain and the relationships between those concepts. It is used to reason about the concepts within that domain. In this paper, we introduce a novel blog categorization method based on leveraging the existing knowledge represented in domain ontology. The novelty of this approach is that it is not dependent on the existence of a training set, as it relies solely on the entities, their relationships, and the taxonomy of categories represented in the ontology. The classification is based on measuring the semantic similarity between the created graph and categories defined in the ontology.

In this section a general overview of the problem is given as well as what motivated us to pursue this problem. The following sections are organized as follows: Section 3 covers a summary of the theoretical description of the work as well as the important terms and definition. Section 4 details data preparation for experiment and design module. Finally, section 5 has discussion of the results and summary, followed by scope for future plans of research work.

3. Theoretical background and definitions:

Ontology: As described by the World Wide Web Consortium (W3C), ontology defines the terms used to describe and represent an area of knowledge. Ontologies are used by people, databases, and applications that need to share domain information (a domain is just a specific subject area of knowledge, such as sports, music, or movies). There are several ways to define ontology. Ontology is defined as formal, explicit specification of a shared conceptualization [8,10]. The goal of ontology is to reduce the conceptual and terminological confusion among the members of a virtual community who need to share documents and information of various kinds [11]. Ontologies have proved their usefulness in different applications scenarios, such as intelligent information integration, knowledge-based systems, natural language processing etc. Supported by ontology, both the user and the system can communicate with each other using a common understanding of a domain [12]. Ontologies are widely used in many areas like AI, biomedical informatics etc.

Regardless of the language used for expression, ontologies share many structural similarities. Common components of the ontologies include:

- Individuals: Instances or objects
- Classes: Sets, collections, concepts, types of objects, or kinds of things. The classes of an ontology may be extensional (characterized only by membership) or intensional in nature.
- Attributes: Aspects, properties, features, characteristics, or parameters that objects can have
- Relations: Ways in which classes and individuals can be related to one another. The most important type of relation is the subsumption relation
- Function terms: Complex structures formed from certain relations that can be used in place of an individual term in a statement
- Restrictions: Formally stated descriptions of what must be true in order for some assertion to be accepted as input
- Rules: Statements in the form of an if-then (antecedent-consequent) sentence that describe the logical inferences that can be drawn from an assertion in a particular form
- Axioms: Assertions (including rules) in a logical form that together comprise the overall theory that the ontology describes in its domain of application. This definition differs from that of "axioms" in generative grammar and formal logic. In these disciplines, axioms include only statements asserted as *a priori* knowledge. As used here, "axioms" also include the theory derived from axiomatic statements.
- Events: The changing of attributes or relations

Ontology created for a given domain hierarchically organized graphical structure which primarily includes a set of concepts as well as relationships connecting them within the domain. Collectively, the concepts and the relationships form a foundation for reasoning about the domain. The classes define the types of attributes, or properties common to

individual objects within the class. Moreover, classes are interconnected by relationships, indicating their semantic interdependence (relationships are also regarded as attributes) [9]. Class hierarchies and class relationships form the schema level of the ontology, while the individuals (object instances or just instances) and links among them (relationship instances) form the so called ground level of the ontology. RDF [14] and OWL are two examples of popular ontology specification languages. A comprehensive, well populated ontology with classes and relationships closely modeling a specific domain represents a vast compendium of knowledge in the domain. Recently, Ontologies have been used in various semantic applications. We believe that the knowledge represented in such a comprehensive ontology can be used to identify topics (concepts) in a text document, provided the document thematically belongs to the domain represented in the ontology. Furthermore, if the concepts in the ontology are organized into hierarchies of higher-level categories, it should be possible to identify the category (or a few categories) that best classify the content of the document.

Three basic components of an OWL ontology that are used in this paper are, Classes, Properties, and instances. Classes are arranged in taxonomy. A subclass inherits all properties of parent and may have its own. Instances of subclass are also instances of superclass. A class may also have more than one super class. This is achieved through using more complex classes with set objects like intersect or union. Ontology properties are tools for expressing the relationship among classes that form a hierarchical superclass subclass relationship.

Tagging: Blogs are widely organized using user provided labels known as blog tags. These textual labels are biased to the user understanding of the domain, and are limited to his linguistic knowledge. Normally a blog tagset contains 0 to 5 labels. Tags behave like meta labels that represents blog content and helps describe an item and allows it to be found again by browsing or searching. Tags are chosen informally and personally by the item's creator or by its viewer. On a website in which many users tag many items, this collection of tags becomes a folksonomy. Tagging was popularized by websites associated with Web 2.0 and is an important feature of many Web 2.0 services. It is now also part of some desktop software. In 2003, the social bookmarking website *Delicious* provided a way for its users to add "tags" to their bookmarks (as a way to help find them later). Delicious also provided browseable aggregated views of the bookmarks of all users featuring a particular tag. This flexibility allows people to classify their collections of items in the way that they find useful, but the personalized variety of terms can make it difficult for people to find comprehensive information about a subject; in order to catch every relevant item, they may have to search several times using different keywords.

TF-IDF : The *tf–idf* weight (term frequency–inverse document frequency) is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the tf–idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.

A high weight in *tf–idf* is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms.

4. Data preparation and processing details

The data for the blog set is collected from the google blog search. More than 1000 blog articles are collected for each of the three domains, sports, music, food.

Blogs are HTML files. A Blog have many features encoded as meta tags, like blog title, author, time stamp, tags, the actual content, that is body of the blog post, comments, and reviews. For the task of classification here only title body and tag set is considered. They need to be preprocessed before the classification task. Here is the description of various modules and their functionalities.

4.1 Algorithm and steps used:

- Preprocessing of the html files for the removal of java scripts CSS. Then blogs are converted to plain text file.
- POS tagging of the text files: Text files are passed to the Stanford POS tagger.
- Stop words removal: Stop words are removed from the extracted parts of speech as they are insignificant for tagging and can skew tf idf scores. There is no definite list of stop words which all natural language processing tools incorporate. Not all NLP tools use a stoplist.
- Stemming: Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form – generally a written word form. We have used tag stemming to improve search

results. The software used to do stemming is ClairLib. The Clair library is a suite of open-source Perl modules. Its architecture also allows for external software to be plugged in with very little effort. Clair lib comprises over 100 modules covering functionality for a wide range of tasks. There are many approaches of extracting the stem of the word, but three major categories are prevalent namely, affix removing, statistical and mixed [13]. The other popular approach is the Porter's stemmer algorithm.

- Term frequency and inverse document frequency calculation: It is calculated as

$$tf \times idf(i,j) = tf(i,j) \times \log\left(\frac{N}{df(j)}\right)$$

Where *tf(i,j)* is the term frequency of term *j* that appears in document i, where *i=1,2,3…,N*. *df(j)* is the document frequency of Term *j* and represents how often Term *j* appears in other documents.

- Finding the words in the blogs that occurred in the ontology before establishing the relationships between those words. The words of the blogs that occurred in the ontology might be entity or instances. These are the nodes of initial semantic graph.

- Semantic graph construction using the ontology and the words that occurred in the blogs.

  Create Matrix: The representative words of the blogs that are represented by the entities and instances in the ontology are selected before creating the matrix. After the selection of the words that occurred in the ontology, the semantic relationships between them is established. If the entity is the subclass of any other entity the value 1 is given and if there exists the relationships between two entities the value 3 is given. If it is an instance of a class/subclass it is given highest score of 4. Any property is given a score of .5. In other cases zero is given.
  Semantic Graph: The matrix itself is the representation of the semantic graph.
  Dominant semantic Graph: The matrix is processed using the breadth first search to find the central entity. The weights are used to find a centrality score for the vertices in the graph. The entity which has the highest weight could be considered as the central entity. The weight of that central entity is saved for further processing.

- **Classification of blogs:** The central entity and it's weight for all the ontology are considered. We take the central entity with the highest weight and the ontology to which it belongs is considered as the classifier of that blogs.

4.2 Classification Accuracy

Classification accuracy is measured for various tags in each of these categories of blogs taken for the experiment. The size of the test set is varied over number of articles in the test dataset for measuring accuracy. The result is shown in figure (1). The result indicates that accuracy tends to stabilize for some threshold number of articles in the test data set.

The central concept(s) for a blog, computed for a threshold score, produces a computed tagset for the corresponding article. This computed tagset is represented as vector, which is then compared to the original tagset of the corresponding article. The resulting accuracies and precision are shown in table 1. The result shows that the proposed approach can produce a more descriptive tagset. The confusion matrix obtained is represented in table, where the cases that constitute TP are the one where both, the terms in the enhanced tagset as well as user annotations agree. FP are the cases where user annotated tagset do not agree with the computed one. Similarly, FN are the cases where the computed tagset indicates the inappropriateness of tags but the given user annotation differ, and TN are the cases where both targets agree on inappropriate tags.

Thus the accuracy and the prescision from the confusion matrix are computed as follows.

Accuracy= (tp+tn) /(tp+fp+fn+tn), and Precision=tp/tp+fp.

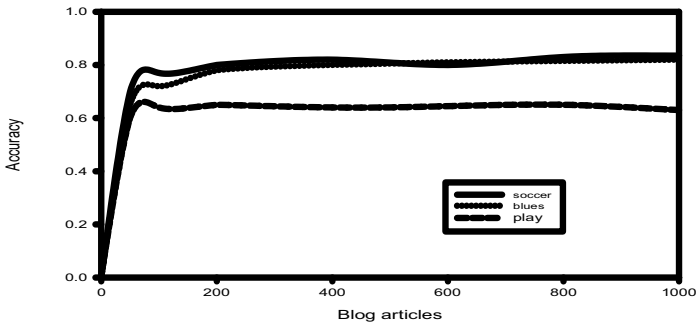

Figure1: Classification accuracy of tag names

	Tf-idf based approach		Ontology based approach	
Blogs	accuracy	precision	accuracy	precision
100	59%	90.9%	64%	94.5%
200	63%	91.8%	71%	94.8%
500	68%	92.6%	78%	92.5%

Table1: Confusion matrix for ontology based and tf-idf based approaches .

5. Results and conclusion

In this paper a blog classification approach using ontology is discussed. The initial results show that this method is effective for weblog classification for the test blog corpus. The comparison with tf-idf based approach shows that the proposed method has improved accuracy and precision.

This paper introduced the use of ontology to conceptually express the meaning of relationships contained in Web documents and suggested an automated document classification method using the ontology. The use of ontology in document classification found out to be effective. The use of ontology has not only classified the blogs to their respective categories but also has established the semantic relationships to understand the underlying concepts the blog is trying to convey.

The presented approach and our experiments indicate that a rich and comprehensive ontology can be successfully used as a text classifier. The selection of a proper mapping between the ontology classes and user defined categories remains as an open question. In the subsequent research it is planned to thoroughly perform the experiment on a wide spectrum of blog corpus, and the fine tuning and optimization of the mechanism for obtaining the centrality score.

References

[1] Xiaochuan Ni, Gui-Rong Xue, Xiao Ling, Yong Yu, Qiang Yang. Exploring in the weblog space by detecting informative and affective articles. In Proc. of WWW '07, pages 281-290, Banff, Alberta, Canada, 2007.

[2] Yun Chi, Shenghuo Zhu, Xiaodan Song, Junichi Tatemura, Belle L. Tseng. Structural and temporal analysis of the blogosphere through community factorization. In Proc. of KDD '07, pages:163-172, San Jose, California, USA, 2007.

[3] Yun Chen, Flora S. Tsai, Kap Luk Chan. Blog search and mining in the business domain. In Proc. of DDDM 07, pages 55–60, San Jose, California, 2007.

[4] Meishan Hu, Aixin Sun, Ee-Peng Lim. Comments-oriented blog summarization by sentence extraction. In Proc. of CIKM '07, pages 901-904, San Jose, California, 2007.

[5] Trevino EM. Blogger motivations: Power, pull, and positive feedback. Internet Research 6.0. 2005.

[6] Jean-Yves Delort. Identifying commented passages of documents using implicit hyperlinks. In Proc. of HYPERTEXT '06, pages 89-98, Odense.

[7] L. Zhou and E. Hovy. On the Summarization of Dynamically Introduced Information: Online Discussions and Blogs. In Proc. of AAAI'06 Spring Symposium on Computational Approaches to Analyzing Weblogs, March 2006.

[8] Gruber, T.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5 (1993) 199-220, 1993

[9] Sheth, A.P., Arpinar, I.B., Kashyap, V.: Relationships at the Heart of Semantic Web: Modeling, Discovering, and Exploiting Complex Semantic Relationships. In 'Enhancing the Power of the Internet: Studies in Fuzziness and Soft Computing'. Springer Verlag (2003)

[10] Tom Grumber, "A translation approach to portable ontology specifications", Knowledge Acquisition, Vol. 5, No. 2, pp.199-220, 1993.

[11] Roberto Navigli, Paola Velardi, "Learning Domain Ontologies for Document Warehouse and Dedicated Web Sites", Computational Linguistics, Vol.30, No.2, pp.152-179, 2004.

[12] V. W. Soo, C. Y. Lin, Ontology-based information retrieval in a multiagent system for digital library, in: Proceedings of the 6th International Conference on Artificial Intelligence and Applications, Taiwan, 2001, pp. 241-246.

[13] Bloehdorn, S., Hotho, A.: Text Classification by Boosting Weak Learners based on Terms and Concepts. 4th IEEE International Conference on Data Mining (ICDM'04) (2004)

[14] Brickley, D., Guha, R.V., RDF Vocabulary Description Language 1.0: RDF Schema. In: McBride, B. (ed.): http://www.w3.org/TR/rdf-schema/ (10 Feb 2004)

[15] Patel-Schneider, P.F., Hayes, P., Horrocks, I., OWL Web Ontology Language Semantics and Abstract Syntax. http://www.w3.org/TR/owl-semantics/ (10 Feb 2004)