Using Tag Semantic Network for Keyphrase Extraction in Blogs

Lizhen Qu Christof Müller Iryna Gurevych
Ubiquitous Knowledge Processing Lab
Computer Science Department, Technische Universität Darmstadt, Hochschulstraße 10
D-64289 Darmstadt, Germany
{qu,mueller,gurevych}@tk.informatik.tu-darmstadt.de

ABSTRACT

Folksonomies provide a comfortable way to search and browse the blogosphere. As the tags in the blogosphere are sparse, ambiguous and too general, this paper proposes both a supervised and an unsupervised approach that extract tags from posts using a tag semantic network. We evaluate the two methods on a blog dataset and observe an improvement in F1-measure from 0.23 to 0.50 when compared to the baseline system.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Information filtering.

General Terms: Algorithms, Experimentation, Performance.

Keywords: Blog,Tag Recommendation, SemRank.

1. INTRODUCTION

Folksonomies are "metadata for the masses" to facilitate search and browsing information in the blogosphere. They reflect directly the vocabulary of users and describe more facets of an object from different points of view than traditional ontological classification schemata. Together with information visualization techniques, they give users a better overview and categorization of the data.

However, we find that tags are very sparse in the blogosphere and most of them indicate only general topical information like "music" and "blog". Uncontrolled vocabularies lead to the ambiguity of tags. A tag suggestion system can facilitate the vocabulary convergence in a manner that users are encouraged to select the suggested ones. In addition, important keyphrases extracted from blog posts can compensate for general tags by providing more focused information. Based on these ideas, this paper proposes both supervised and unsupervised ways to overcome the weaknesses of human tagging by using a tag semantic network derived from the existing tag space. This is a different approach as opposed to the existing tag recommendation systems in the blogosphere [4, 7], since tag candidates are extracted directly from the blog content, thus the existing tag space is expanded with novel and more focused tags. Also, the state-of-the-art keyphrase extraction systems like [3, 8, 9] are tuned for keyphrases assigned by professional indexers, we show in our evaluation that two of them can only achieve low performance on the blog data.

2. TAG SEMANTIC NETWORKS

A tag semantic network consists of the nodes representing individual tags and the edges representing the strength of semantic relatedness. Assuming that similar tags are used to annotate similar blog posts, the latter provide the context for the tag usage. We use the mean tf.idf vector of the associated blog posts to model the context of a given tag t, which is defined as

$$context(t) = \frac{1}{n} \sum_{d \in D_t} tfidf_{norm}(d)$$

where D_t is the collection of posts that the tag t is attached to and n is the size of the collection. The function $tfidf_{norm}(d)$ calculates the normalized tf.idf vector of a document d, so that the Euclidean norm of the vector is 1.

We observe that web users tend to annotate one post with several tags of different abstraction levels like "jazz" and "music". The semantic relatedness between them can be expressed with an association rule $jazz \Rightarrow music$. We use *confidence* $P(Y \mid X) = P(X,Y)/P(X)$ [1] to define the rule strength between two tags X and Y. In our supervised learning method, we also use the undirectional measure *interest* $P(A,B)/(P(A) * P(B))$ [2] to characterize the co-occurrence between the two tags A and B.

3. A SUPERVISED APPROACH

Following a supervised learning approach, our keyphrase extraction process is separated into **candidate phrase identification** and **keyphrase filtering**.

Candidate phrases are all unigrams, bigrams and trigrams having predefined noun based POS patterns[1]. These phrases cover already 90% of tags used as gold standard.

The **filtering process** utilizes a learned model to identify the most important tag candidates among all selected n-grams. The features not depending on the tag semantic networks are: i. tf.idf; ii. first occurrence; iii. word length of the n-grams; iv. named entity class of n-grams; v. whether the n-grams occur within a html hyperlink; vi. whether the n-grams occur within emphasizing html markups. By utilizing the derived tag networks, *statistical phrase relatedness* is the sum of context similarity and the *interest* measure to other candidate phrases identified in the first step. The context similarity of two terms is defined as the cosine distance of the two corresponding context vectors. It is considered only if it is over a certain threshold (0.2 in our experiments).

[1] The patterns are *N N*, *A N*, *N*, *N U N*, *A U N*, *A Prep N* and *N Prep N*, as identified by Minipar [5].

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10.

The logistic regression classifier from Autonlab[2] is applied to the feature vectors of candidate terms to identify the top 5 ranked candidate terms as the final set of keyphrase.

4. AN UNSUPERVISED APPROACH

Like TextRank[9], our SemRank algorithm considers each document as a directed text graph $G(V, E)$, where V denotes the set of vertices representing lemmas of occurring noun and adjective unigrams, E denotes the set of weighted edges. As in [9], $In(V_i)$ denotes the collection of vertices pointing to V_i and $Out(V_i)$ is the set of vertices pointed by V_i.

We define different edge weights in our experiments. i. $w_{ij}^{cooc} = \frac{1}{\#Out(V_j)}$ [9] measures the co-occurrence of lemmas within a window of size N. ii. If two lemmas i and j are matched in the tag network, the weight of edges w_{ij}^{conf} is set to their $confidence$. iii. w_{ij}^{cont} defines the cosine distance between the two corresponding mean tf.idf vectors. iv. if two lemmas i and j are found as Hypernym or Meronym within a distance of K in WordNet, w_{ij} is set to $\frac{1}{1+K}$.

Let w_{ij} be the weight from vertex j to i, we use PageRank [6] to calculate the ranking score $S(V_i)$ of V_i:

$$S_{t+1}(V_i) = (1-d) + d \times \sum_{V_j \in In(V_i)} \frac{w_{ij}}{\sum_{V_k \in Out(V_j)} w_{kj}} S_t(V_j)$$

where the dumping factor d is set to 0.85 in our expriments. When stacking all $S(V_i)$ into a vector s, we get the following formula to represent the ranking scores of lemmas at iteration $t + 1$.

$$s_{t+1} = (1-d)p + dMs_t$$

Here $p = [1, ..., 1]^T$ is the preference vector and $M = \{m_{ij}\}$ is the transition matrix with $m_{ij} = \frac{w_{ij}}{\sum_{V_k \in Out(V_j)} w_{kj}}$.

According to PageRank, a random surfer starts a new navigation by jumping to a page picked uniformly and randomly from the collection. Since each lemma is of different significance, we use a normalized tf.idf vector to replace the original uniform distribution p to model the “preference”.

By using the matrix interpretation, the different types of edge weight definition can also be used together, which can be seen as the sum of the corresponding transition matrices. To ensure the convergence of the algorithm, the sum of the matrices is normalized so that the Manhattan norm of each column is equal to 1.

After ranking, all unigrams with the top 5 scores are selected as candidates. These candidates, if located adjacent to each other, are collapsed into new multi-word keyphrases.

5. EVALUATION AND ANALYSIS

In our experiments, we use 621 English posts with at least 3 gold standard tags selected from the ICWSM 2006[3] blog corpus for evaluation and 2500 for building the tag semantic network. The tags are chosen as the gold standard, if they occur at least once in the posts. To evaluate our supervised method, 434 of them are used for training and 187 for testing. On the same test dataset, the KEA system [8], TextRank and SemRank are evaluated for comparison.

The evaluation measure we use is based on the exact stem matching of tags. The results are given in terms of precision, recall and F1-measure.

	Recall	Precision	F-measure
KEA	0.24	0.22	0.23
Supervised	**0.52**	0.48	**0.50**
TextRank	0.19	0.24	0.21
SemRank w^{cooc} WordNet	0.19	0.07	0.11
SemRank tf.idf w^{cont} w^{conf}	0.49	**0.50**	0.49

Table 1: Results of keyphrase extraction.

Table 1 shows the results of our experiments, where w^X of SemRank denotes the used edge weight definitions and tf.idf denotes that the preference vector is replaced by the tf.idf vector. Our supervised and unsupervised methods outperform the baselines significantly. Both methods achieve over 200% improvement when using the information from the tag semantic network. The experiment using WordNet is a quantitative evidence that folksonomies indeed capture different relations and facets of an object or event than traditional ontologies. Furthermore, our empirical results show that the preference vector provides an effective way to improve PageRank like algorithms by integrating the information like tf.idf. We also observe that over 70% of tags in our datasets are single words scattered over multiple sentences, while the keyphrases in the test dataset used in [3, 9] are mainly multi-word phrases. This is an explanation why the POS pattern features in [3] and the post-processing step in [9] cannot achieve the large improvement as in the original datasets used for evaluation.

6. ACKNOWLEDGEMENT

This work has been partly supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806 and by the German Research Foundation under grant No. GU 798/1-3.

7. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. *Proc. of the 1993 ACM SIGMOD International Conf. on Management of Data*, pages 207-216. ACM, 1993.

[2] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: generalizing association rules to correlations. *Proc. of the 1997 ACM SIGMOD International Conf. on Management of Data*, pages 265-276, 1997.

[3] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. *Proc. of the 2003 Conf. on Empirical Methods in NLP*, pages 216-223. Association for Computational Linguistics, 2003.

[4] R. Jaschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in folksonomies. Knowledge Discovery in Databases: PKDD 2007, 4702:506-514, 2007.

[5] D. Lin. Dependency-based evaluation of MINIPAR. *Proc. of the Workshop on Evaluation of Parsing Systems*, 1998.

[6] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[7] Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: Collaborative tag suggestions. *WWW2006: Proc. of the Collaborative Web Tagging Workshop* 2006.

[8] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning. KEA: practical automatic keyphrase extraction. *DL 99: Proc. of the fourth ACM conf. on Digital libraries*, pages 254-255. ACM, 1999.

[9] R. Mihalcea and P. Tarau. TextRank: bringing order into texts. *vertex*, 4:6, 2006.

[2] http://www.autonlab.org

[3] http://www.icwsm.org/data.html