Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Wikitop: Using Wikipedia Category Network to Generate Topic Trees

Saravana Kumar, Prasath Rengarajan, and **Arockia Xavier Annie**
Department of Computer Science and Engineering, College of Engineering, Guindy
Anna University, Sardar Patel Road, Guindy, Chennai, Tamil Nadu 600025
{savkumar90,prasathindiarajan,aucs.annie}@gmail.com

Abstract

Automated topic identification is an essential component in various information retrieval and knowledge representation tasks such as automated summary generation, categorization search and document indexing. In this paper, we present the Wikitop system to automatically generate topic trees from the input text by performing hierarchical classification using the Wikipedia Category Network (WCN). Our preliminary results over a collection of 125 articles are encouraging and show potential of a robust methodology for automated topic tree generation.

Introduction

Automated topic identification (ATI) of textual data is the process of assigning a collection of tags that best represent the input text. ATI can also be seen as a text categorization or classification task. Currently prevalent research in ATI is focused on generating a *flat* structure of tags by semantic analysis methods such as word frequency counting and title-keyword techniques. However, a flat structure may not be a sufficient representation of the input text as it omits conceptual generalisations. Moreover, as shown in (Coursey and Mihalcea 2009), the topic of a text may not always be apparent in the input text and may need an external knowledge source to help predict the topic. Topic trees or category trees are a collection of tags in tree form with the tags organised in descending order of generality. The tag representing the most general concept occupies the root of the tree and the tags representing the least general concepts occupy the leaves of the trees.

The Wikipedia Category Network (WCN) is the logical organisation of all Wikipedia categories, subcategories and pages in the form of a tree. Although the category structure of Wikipedia is a directed cyclic graph, we develop the WCN as a tree by ignoring cycle causing edges and develop a topic hierarchy.

We build upon prior research in topic tree generation (Sun and Lim 2001; Lin 1995) and hierarchical classification (Koller and Sahami 1997) to develop the Wikitop system for automated topic tree generation by hierarchical classification using the Wikipedia Category Network as a topic hierarchy. We formally define topic trees and topic hierarchies in section .

We present a supervised classification approach using Support Vector Machines that exploits the category structure of Wikipedia and uses it as an initial topic hierarchy to perform hierarchical classification and subsequently build topic trees. Moreover, since Wikipedia is the largest curated online encyclopaedia, there are an enormous number of articles labelled with their categories. Therefore, there is a massive amount of training data. We evaluate an initial prototype of the Wikitop system by passing a collection of 125 articles to prove its feasibility.

Automated Topic Tree Generation

A *topic hierarchy* can be defined as topic tags organised in tree structure, where every parent node is a conceptual generalisation of its children.

For a given input text document, the problem of automatic topic recognition involves assigning topic tags from the topic hierarchy to an output *topic tree* such that, the tags are the best description of the topic of the text and are organised in the same way as they are in the topic hierarchy where each parent is a conceptual generalisation of it's children

Methodology

We adapt the concept of hierarchical classification from (Koller and Sahami 1997) and use the Wikipedia Category Network (WCN) as a topic hierarchy. A multi-class multi-label classifier is trained at every node of the WCN. The classifier at each node uses the names of the categories directly connected and one level below it as classes for training. We use multi-class multi-label classifiers because at each level, there maybe more than two categories directly connected to it and a given text can belong to more than one of those categories. For training samples each class is assigned the pages that are directly connected to both itself and to categories that come under it up to a depth of δ. The δ values are an indication of the number of training example. If $\delta = 0$ only pages that are directly connected to the class node are considered as training example for that class. If $\delta = 1$ the pages that are directly connected to the class node and pages that belong to the category immediately connected to the class node are considered for training and so

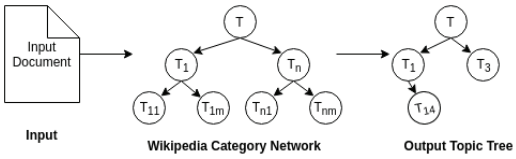

Figure 1: Architecture of the Wikitop system.

on. With an increase in δ the number of training example for each classifier increases.

In order to build the topic tree, a root node for the topic tree is first chosen from the topic hierarchy. The classifier for the chosen node from the WCN is selected and the input text is passed through the classifier. The input text is tokenised into sentences before it is passed through the classifier. For each sentence, the classifier returns probabilities for each class. The probability for a class has to be above a certain decision boundary Φ to consider the class as a prediction for a particular sentence. The Φ value is an indication of how wide the topic tree must be. Smaller values of Φ lead to wider trees as classes with smaller predicted probability values for a sentence are also considered in the final output.

The distinct set of classes that are predicted across all sentences are added as children to the root node in our topic tree. Each of these children are then considered as a root and this process is repeated for σ levels. The parameter σ is the depth of our topic tree. (i.e) when $\sigma = 3$ a topic tree of depth 3 is generated.The σ value indicates the generalisation of the topic tree. Higher values of σ lead to the construction of highly specific topic trees and lower values of σ lead to more general topic trees.

To perform preliminary analysis to test our hypothesis and methodology, Support Vector Machine(SVM) classifiers were used based on the reasoning found in (Joachims 1998) with a Linear Kernel and balanced class weights.

Evaluation

In order to test the feasibility of the Wikitop system and present a preliminary evaluation, 125 candidate articles were first identified from the WCN. The criteria used to select those articles was that the articles must not directly belong to any category that is within a depth of 5 from the category titled 'Computer Science'. This condition was necessary because testing was done with a maximum $\delta = 2$ and $\sigma = 3$. This means that any page that is not connected to a node within a depth of 5 from the root node would not have been a training example for any of the classifiers. Subsequently, the node 'Computer Science' was set as the root node and the articles were passed through the WCN. Different values of δ and Φ were used to evaluate the accuracy metrics for our model. The article along with the output topic trees were shown to independent annotators who were asked if they agreed with the topic classification or not. Any article with more than 50% of the tags classified correctly was considered as partially accurate. The results of the evaluation are summarised in 1.

Conclusion

The results in 1 indicate an increase in accuracy with an increased value of δ. By increasing the δ value, there is an increase in the number of training examples for each classifier and therefore there is a better overall accuracy for the Wikitop system. However, with increased values of δ the complexity of training the classifier nodes in the WCN is also higher.

The preliminary results indicate that the Wikitop system is a robust method for topic tree generation. It is important to note that the WCN was trained with only one type of classifier (SVM) with default parameters. This work in progress will explore using different types of classifiers and optimising those classifiers.

Table 1: It can be seen that by increasing the δ value , there is an increase in the accuracy.

δ	Φ	**Partial Accuracy**	**Accuracy**
0	0.1	25.60±2%	21.20±2%
0	0.5	17.60±2%	16.00±2%
1	0.1	77.50±2%	60.62±2%
1	0.5	79.46±2%	75.00±2%
2	0.1	96.57±2%	85.00±2%
2	0.5	76.38±2%	72.91±2%

Acknowledgements

We would like to thank Muthu Palaniappan Alagappan, Srideepika Jayaraman, Makesh Narsimhan and Akshay Ganapathy for their contribution to this work.

References

Coursey, K., and Mihalcea, R. 2009. Topic identification using wikipedia graph centrality. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, 117–120. Association for Computational Linguistics.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, 137–142. Springer.

Koller, D., and Sahami, M. 1997. Hierarchically classifying documents using very few words.

Lin, C.-Y. 1995. Knowledge-based automatic topic identification. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, 308–310. Association for Computational Linguistics.

Sun, A., and Lim, E.-P. 2001. Hierarchical text classification and evaluation. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, 521–528. IEEE.