System for semi-automatic ontology construction

Blaž Fortuna
Blaz.Fortuna@ijs.si
Institute Jožef Stefan
Jamova 39
1000, Ljubljana

Marko Grobelnik
Marko.Grobelnik@ijs.si
Institute Jožef Stefan
Jamova 39
1000, Ljubljana

Dunja Mladenič
Dunja.Mladenic@ijs.si
Institute Jožef Stefan
Jamova 39
1000, Ljubljana

ABSTRACT

In this paper, we review two techniques for topic discovery in collections of text documents (Latent Semantic Indexing and K-Means clustering) and present how we integrated them into a system for semiautomatic topic ontology construction. The system offers supports to the user during the construction process by suggesting topics and analyzing them in real time.

General Terms

Algorithms, Human Factors.

Keywords

Semi-automatic Ontology construction, Background knowledge

1. INTRODUCTION

When working with large corpora of documents it is hard to comprehend and process all the information contained in them. Standard text mining and information retrieval techniques usually rely on word matching and do not take into account the structure of the documents within the corpus. We try to overcome that by automatically extracting the topics covered within the documents from the corpus and helping the user to organize them into a topic ontology.

Topic ontology is a set of topics connected with different types of relations. Each topic includes a set of related documents. Construction of such ontology from a given corpus can be a very time consuming task for the user. In order to get a feeling on what the topics in the corpus are, what the relations between topics are and to assign each document to some certain topics, the user has to go through all the documents and process them manually. We tried to overcome this by building *OntoGen*, a special tool which helps the user by suggesting the possible new topics and visualizing the topic ontology created so far, all in real time. OntoGen, in combination with the corpus visualization tools [4], aims at assisting the user in a fast semi-automatic construction of the topic ontology from a large document collection.

This paper is organized as follows. In Section 2 we present text mining techniques that are used in OntoGen, and in Section 3 we give a short demonstration of the tool and its features.

2. TEXT MINING TECHNIQUES

2.1 Representation of text documents

In order to use the algorithms we will describe later we must first represent text documents as vectors. We use standard Bag-of-Words (BOW) approach together with the TFIDF weighting [5]. This representation is often referred to as vector-space model. The similarity between two documents is defined as the cosine of the angle between their vector representations – cosine similarity.

2.2 Latent Semantic Indexing

The language contains much redundant information, since many words share common or similar meaning. For computer this can be difficult to handle without some additional information (background knowledge). Latent Semantic Indexing (LSI), [3], is a technique for extracting this background knowledge from text documents. It uses a technique from linear algebra called Singular Value Decomposition (SVD) and bag-of-words representation of text documents for detecting words with similar meanings. This can also be viewed as extraction of hidden semantic concepts or topics from the text documents.

2.3 K-Means clustering

Clustering is a technique for partitioning data so that each partition (or cluster) contains only points which are similar according to some predefined metric. In the case of text this can be seen as finding groups of similar documents, that is documents which share similar words.

K-Means [6] is an iterative algorithm which partitions the data into k clusters. It has already been successfully used on text documents [7] to cluster a large document corpus based on the document topic.

2.4 Keywords extraction

We used two methods for extracting keywords from a given set of documents: (1) keyword extraction using centroid vectors and (2) keyword extraction using Support Vector Machine (SVM) [2]. We used this two methods to generate description for a given topic based on the documents inside the topic.

The first method works by using the centroid vector of the topic (centroid is the sum of all the vectors of the document inside the topic). The main keywords are selected to be the words with the highest weights in the centroid vector. The second method is based on the idea presented in [1] which uses SVM binary classifier. Let A be the topic which we want to describe with keywords. We take all the documents from the topics that have A for a subtopic and mark these documents as negative. We take all the documents from the topic A and mark them as positive. If one document is assigned both negative and positive label we say it is positive. Then we learn a linear SVM classifiers on these documents and classify the centroid of the topic A. Keywords describing the concept A are the words, which's weights in SVM normal vector contribute most when deciding if centroid is positive.

The difference between these two approaches is that the second approach takes into account the context of the topic. Let's say that we have a topic named 'computers'. When deciding, what the keywords for some subtopic A are, the first method would only look at what the most important words within the subtopic A are and words like 'computer' would most probably be found important. However, we already know that A is a subtopic of 'computers' and we are more interested in finding the keywords

that separate it from the other documents within the 'computers' topic. The second method does that by taking the documents from all the super-topics of A as a context and learns the most crucial words using SVM.

3. SEMI-AUTOMATIC CONSTRUCTION OF TOPIC ONTOLOGY

We view semi-automatic topic ontology construction as a process where the user is taking all the decisions while the computer only gives suggestions for the topics, helps by automatically assigning documents to the topics, helps by suggesting names for the topics, etc. The suggestions are applied only when the users decides so. The computer also helps by visualizing the topic ontology and the documents.

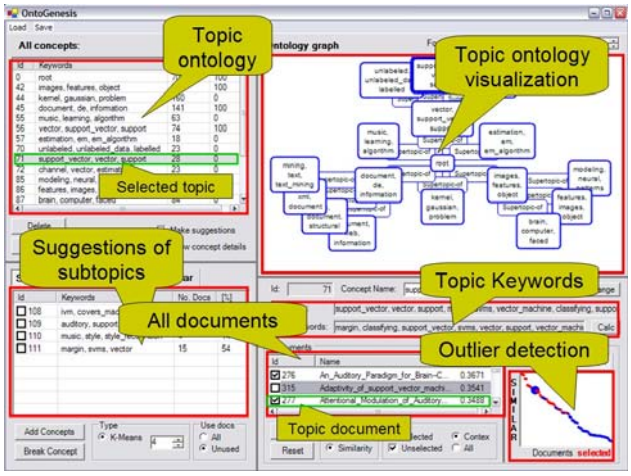

Figure 1. Screen shot of the interactive system OntoGen for construction of topic ontologies.

In Figure 1 you can see the main window of the interactive system we developed. The system has three major parts that will be further discussed in following subsections. In the central part of the main window is a visualization of the current topic ontology (Ontology visualization). On the left side of the window is a list of all the topics from this ontology. Here the user can select the topic he wants to edit or further expand into subtopics. Further down is the list of suggested subtopics for the selected topic (Topic suggestion) and the list with all topics that are in relation-ship with the selected topic. At the bottom side of the window is the place where the user can fine-tune the selected topic (Topic management).

3.1 Ontology visualization

While the user is constructing/changing topic ontology, the system visualizes it in real time as a graph with topics as nodes and relations between topics as edges. See Figure 1 for an example of the visualization.

3.2 Topic suggestion

When the user selects a topic, the system automatically suggests what the subtopics of the selected topic could be. This is done by LSI or k-means algorithms applied only to the documents from the selected topic. The number of suggested topics is supervised by the user. Then, the user selects the subtopics he finds reasonable and the system automatically adds them to the ontology with relation 'subtopic-of' to the selected topic. User can also decide to replace the selected topic with the suggested subtopics. In Figure 1 you can see how is this feature implemented in our system.

3.3 Topic management

The user can manually edit each of the topics he added to the topic ontology. He can change which documents are assigned to this topic (one document can belong to more topics), what is the name of the topic and what is the relationship of the topic to other topics. The main relationship is subtopic-of and is automatically added when adding subtopics as described in the previous section. The user can control all the relations between topics by adding, removing, directing and naming the relations.

Here the system can provide help on more levels:

- The system automatically assigns the documents to a topic when it is added to the ontology.
- The system helps by providing the keywords describing the topic using the methods described in Section 3. This can assist user when naming the topic.
- The system computes the cosine similarity between each document from the corpus and the centroid of the topic. This information can assist the user when searching for documents related to the topic. The similarity is shown on the list of documents next to the document name and the graph of similarities is plotted next to the list. This can be very practical when searching for outliers inside the concepts or for the documents that are not in the concepts but should be in considering their content.

REFERENCES

[1] Brank J., Grobelnik M., Milic-Frayling N. & Mladenic D. Feature selection using support vector machines. Proc. of the Third International Conference on Data Mining Methods and Databases for Engineering, Finance, and Other Fields, 2002.

[2] Cristianini N. & Shawe-Taylor, J., An introduction to support vector machines, Cambridge University Press

[3] Deerwester S., Dumais S., Furnas G., Landuer T. & Harshman R. Indexing by Latent Semantic Analysis, J. of the American Society of Information Science, vol. 41/6, 391-407

[4] B. Fortuna, D.Mladenic, M. Grobelnik. Visualization of text document corpus. Informatica journal 29 (2005), 497-502

[5] Grobelnik M. & Mladenic D. Automated knowledge discovery in advanced knowledge management. J. of. Knowledge management 2005, Vol. 9, 132-149.

[6] Jain, A. K., Murty M. N., & Flynn P. J. Data Clustering: A Review, ACM Computing Surveys, vol 31/3, 264-323, 1999.

[7] Steinbach, M., Karypis, G., Kumar, V. A comparison of document clustering techniques. In Proceedings of KDD Workshop on Text Mining, pp. 109110, 2000