SEMANTIC DOCUMENT CLASSIFICATION

Nihan Gümüş-Merve Uğur-Mustafa Güngör
Advisor:Erdoğan Doğdu

Çankaya University, Department of Computer Engineering

Abstract

The web is big and there are billions of text documents that cannot be searched efficiently. Most of the current efforts depend of the word vector representation of text documents, and since this does not involve any semantics, the classification methods are not very accurate. In this project we will investigate existing or currently being developed novel methods in semantic classification of text documents, test and extend them, and also develop new semantic methods to classify documents more accurately and compare all methods extensively.

Introduction

Grouping similar features means that the information obtained is valid for that group. The process of partitioning a set of data objects into subsets is Clustering. Document Clustering has wide application fields. Document clustering is used for data mining, information retrieval, and knowledge discovery from a data of different category. Document clustering is an application that allows you to perform cluster analysis on text documents. Document clustering is an application that allows you to perform cluster analysis on text documents. The document clustering process includes and uses identifiers and conventions of these identifiers. The identifiers of the words in the cluster are recognized.

Solution

We have more than one document. Today, text documents; the internet, e-mail and electronic data base a lot on web pages and are stored in the format. Organize and browse through the papers every day. To overcome this problem are being used more than one method. This paper discusses the methods used. The purpose of a common set of semantic document set that contains a single semantic web documents logically is to categorize the document set. Understanding automatic and smart text due to insufficient billions of text that cannot be efficiently searching document. Vector representation of most current efforts depends on the text document and this does not include any semantics, classification methods are not very accurate. Classification of text documents in this project or still present in developing new semantic method are investigated and dialectics
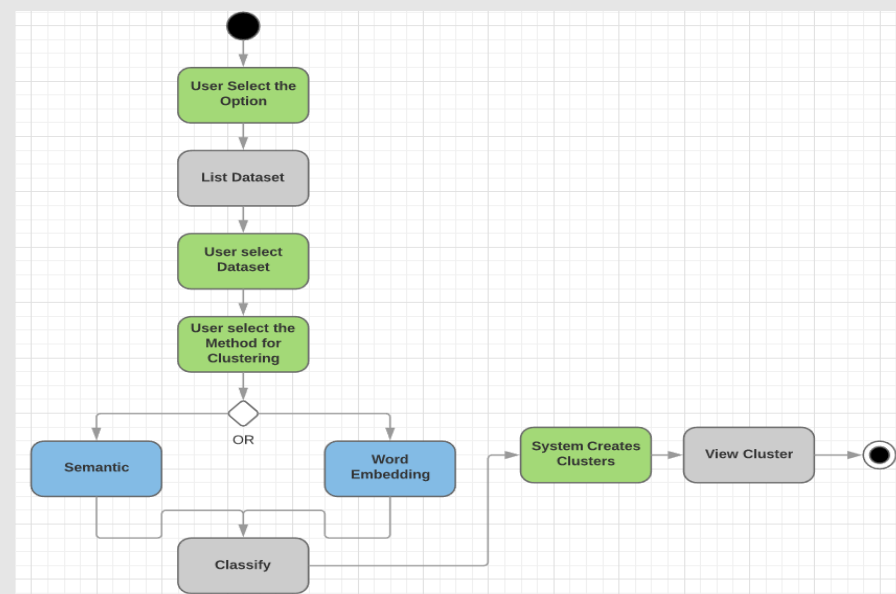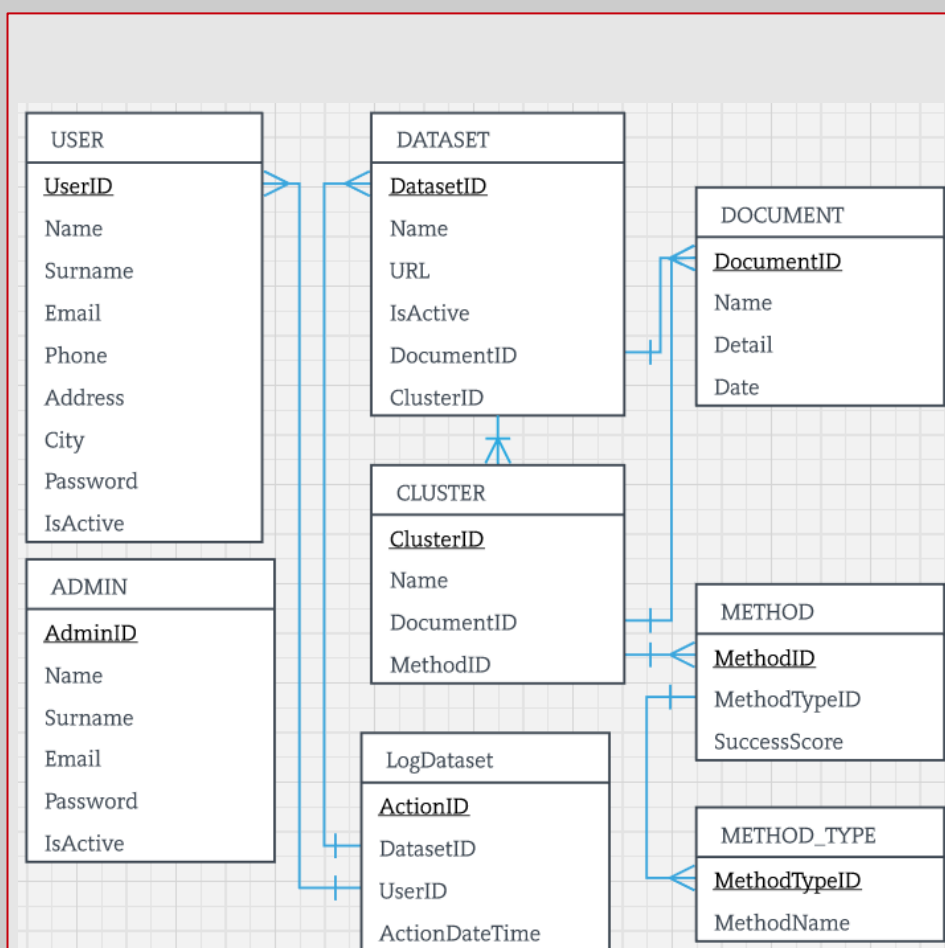

Figure 1 - Flowchart
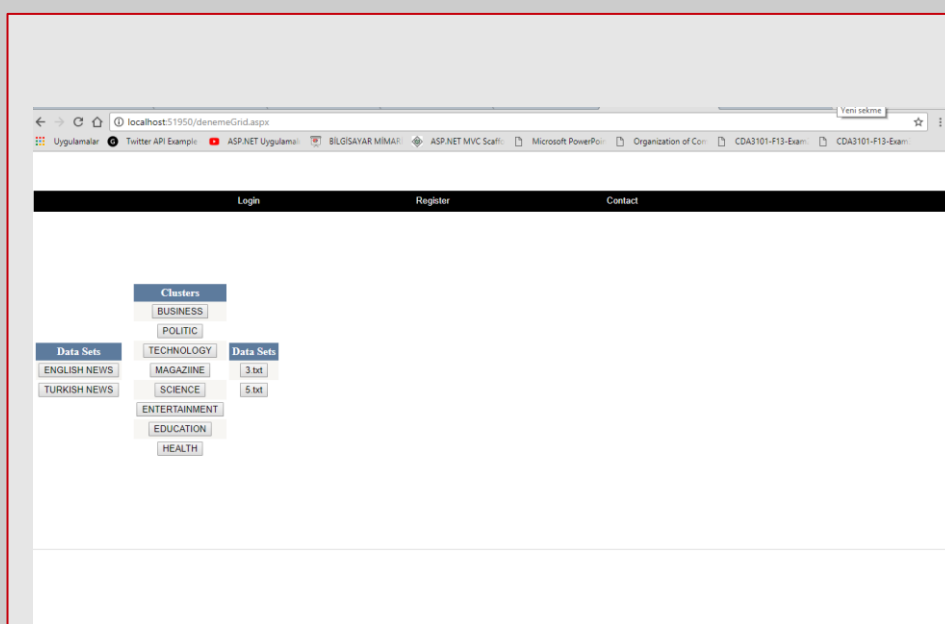

Figure 2 – Database

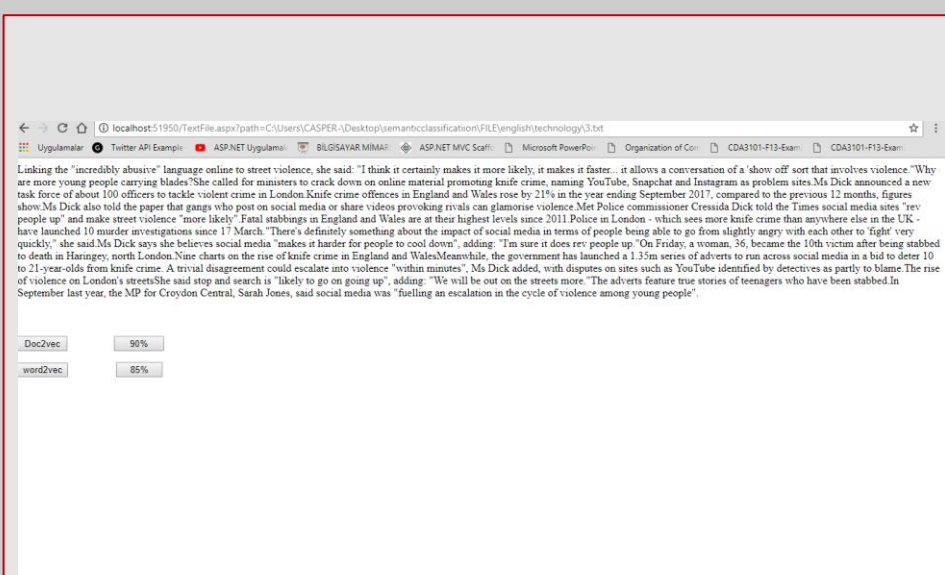**Figure 3 – Finished Product**

Figure 4 – Finished Product

Results & Conclusion

We have more than one document. Today, text documents; the internet, e-mail and electronic data base a lot on web pages and are stored in the format. Organize and browse through the papers every day. To overcome this problem are being used more than one method. This paper discusses the methods used. The purpose of a common set of semantic document set that contains a single semantic web documents logically is to categorize the document set. Understanding automatic and smart text due to insufficient billions of text that cannot be efficiently searching document. Vector representation of most current efforts depends on the text document and this does not include any semantics, classification methods are not very accurate. Classification of text documents in this project or still present in developing new semantic method are investigated and dialectics.

Acknowledgement