Abstract

Now a day's detection of templates from a large number of web pages has received a lot of attention. Template detection technique improves the performance of clustering, classification & search engines. In our work we proposed a novel algorithm by using cosine similarity based Template Extraction. We are using the cosine similarity approach to cluster the web documents. With the help of underlying structure of web documents we found the template for individual cluster. Our experimental evaluation show that our approach is effective in terms of computing Time and Clustering cost.

References

- M. de Castro Reis, P. B. Golgher, A. S. da Silva, and A. H. F. Laender,
- Chulyun Kim and Kyuseok Shim, Member, IEEE "TEXT: Automatic Template Extraction from Heterogeneous Web Pages";

Index Terms

Computer Science

Information Sciences

Keywords

Template Extraction TEXT_MDL TEXT_MAX Cosine similarity clustering.