Abstract

The study of identification of near duplicate content involves identifying search categories which generate same URL in a query result. These categories are needed to be identified so that results can be improved by removing duplicate URL. Generating same URL in results irritates the user and it also decreases priority of other URL. These URL displayed on second or third page which user do not bother to open. Near duplicate content sometimes hides better results from the user and make the search results ineffective. There are many algorithms and procedures or filters to reduce the duplicity. But to reduce duplicity there is need to identify that duplicates. Which categories generate most duplicate results, in what form redundancy exists, which search engine generates these duplicate results and so on. This paper shows efforts to identify categories with maximum duplicates in term of same URL.

References


Index Terms

Computer Science  Information Sciences

Keywords

Keywords are your own designated keywords which can be used for easy location of the manuscript using any search engines.