In today’s world internet is the answer to every question. So at any time one can easily copy the content from web and use it. This is known as plagiarism. It is growing now days. Usually in plagiarism people reword the documents, copy them, do not give references. It is difficult to detect plagiarism as people rephrase the text do not copy it directly. To detect plagiarism apache lucene have been used. Firstly indexing of the original document is done and then used cosine similarity to compare the plagiarised document with set of documents which are there saved previously.

References

University, June 2011
5. Bela Gipp Norman Meuschke, "Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence", Mountain View, CA, USA, September 2011

Index Terms
Computer Science Pattern Recognition

Keywords
Apache Lucene, Indexing, Cosine similarity, Plagiarism