Correlation-based Document Clustering using Web Logs*

Zhong Su1, Qiang Yang2, Hongjiang Zhang3, Xiaowei Xu4, Yuhen Hu5
1Department of Computing Science, Tsinghua University, Beijing 100084, China
2 School of Computing Science, Simon Fraser University, Burnaby, BC Canada V5A 1S6
3 Microsoft Research China, 5F, Beijing Sigma Center, Beijing 100080 P.R. China
4 Siemens AG, Information and Communications Corporate Technology, D-81730 Munich, Germany
5 Department of E&E, University of Wisconsin-Madison, Madison, WI 53706 USA

suzhong bj@hotmail.com1, qyang@cs.sfu.ca2, hjzhang@microsoft.com3,
xiaowei.xu@mchp.siemens.de4, Hu@engr.wisc.edu5

Abstract
A problem facing information retrieval on the web is how to effectively cluster large amounts of web documents. One approach is to cluster the documents based on information provided only by users' usage logs and not by the content of the documents. A major advantage of this approach is that the relevancy information is objectively reflected by the usage logs; frequent simultaneous visits to two seemingly unrelated documents should indicate that they are in fact closely related. In this paper, we present a recursive density based clustering algorithm that can adaptively change its parameters intelligently. Our clustering algorithm RDBC (Recursive Density Based Clustering algorithm) is based on DBSCAN, a density based algorithm that has been proven in its ability in processing very large datasets. The fact that DBSCAN does not require the pre-determination of the number of clusters and is linear in time complexity makes it particularly attractive in web page clustering. It can be shown that RDBC require the same time complexity as that of the DBSCAN algorithm. In addition, we prove both analytically and experimentally that our method yields clustering results that are superior to that of DBSCAN.

1. Introduction

A problem facing information retrieval on the web is how to effectively cluster large amounts of web documents. One approach is to cluster the documents based on information provided only by users' usage logs and not by the content of the documents. A major advantage of this approach is that the relevancy information is objectively reflected by the usage logs; frequent simultaneous visits to two seemingly unrelated documents should indicate that they are in fact closely related. In this paper, we present an efficient algorithm for clustering large sets of web documents based on distance measures that are provided by only the server log data.

There is a great deal of work done previously in clustering, including K-means [6], HAC[3][12][1], CLANRNS [11] etc. In the IR community, the Scatter/Gather algorithm [5] is aimed at re-organizing document search results by examining document contents. It is similar to K-means in that it requires pre-set cluster number, which is a requirement that we do not assume in our paper. Suffix-Tree [14] is another closely related clustering method. Its input is also portions of the document contents and thus is different from the problem we face.

Because we only have server log information, we can build a distance metric similar to that by [9]. Based on this distance information, we choose to extend DBSCAN [7], an algorithm to group neighboring objects of the database into clusters based on local distance information. It is very efficient because only one scan through the database is required. Moreover, it does not require a predetermined cluster number to operate.

DBSCAN constructs clusters using distance transitivity based on a density measure defined by the user. Documents that have many co-visited documents around them are considered dense. DBSCAN performs this clustering using a fixed threshold value to determine “dense” regions in the document space. Because this threshold value is constant across all points in the space, the algorithm often cannot distinguish between dense and loose points, and as a consequence, often the entire document space is lumped into a single cluster.

* This work was performed in Microsoft Research China