

NETWORK THEORY AND APPLICATIONS

Clustering and Information Retrieval

Weili Wu, Hui Xiong and Shashi Shekhar
Editors

Kluwer Academic Publishers

Clustering in Metric Spaces with Applications to Information Retrieval

Ricardo Baeza-Yates

Benjamín Bustos

Center for Web Research, Dept. of Computer Science

Universidad de Chile, Blanco Encalada 2120, Santiago, Chile

E-mail: {rbaeza,bbustos}@dcc.uchile.cl

Edgar Chávez

Universidad Michoacana, Morelia, México

E-mail: elchavez@fismat.umich.mx

Norma Herrera

Univ. Nacional de San Luis, San Luis, Argentina

E-mail: nherrera@unsl.edu.ar

Gonzalo Navarro

Center for Web Research, Dept. of Computer Science

Universidad de Chile, Blanco Encalada 2120; Santiago, Chile

E-mail: gnavarro@dcc.uchile.cl

Contents

1	Introduction	2
2	Our Clustering Method	4
2.1	Clustering in Metric Spaces	6
2.2	Mutual k-Nearest Neighbor Graph Clustering Algorithm	7
2.2.1	The Clustering Algorithm	9
2.2.2	Connectivity Properties	10
2.3	The Range r Graph	11
2.3.1	Outliers, Equivalence Classes and Stability	12

2.4	Radius vs. Neighbors	13
2.5	The Connectivity Parameters	14
2.6	Intrinsic Dimension	15
3	Morphological Stemming and the Holomorphic Distance	15
3.1	Motivation	15
3.2	The Holomorphic Transformation	17
3.3	A Morphological Stemmer Using Clustering	17
4	Clustering for Approximate Proximity Search	18
4.1	The Vector Model for Information Retrieval	21
4.2	Techniques for Approximate Proximity Searching	22
4.3	Experimental Results	24
5	Clustering for Metric Index Boosting	24
5.1	GNATs	27
5.2	Experimental Analysis	29

References