

Graph-Theoretic Techniques for Web Content Mining

Adam Schenker

University of South Florida, USA

Horst Bunke

University of Bern, Switzerland

Mark Last

Ben-Gurion University of the Negev, Israel

Abraham Kandel

University of South Florida, USA

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRUNG TÂM THÔNG TIN THƯ VIỆN

A-00/ 5216

 **World Scientific**

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Contents

<i>Preface</i>	vii
1. Introduction to Web Mining	1
1.1 Overview of Web Mining Methodologies	3
1.2 Traditional Information Retrieval Techniques	5
1.2.1 Vector-based distance measures	6
1.2.2 Special considerations for web documents	7
1.3 Overview of Remaining Chapters	9
2. Graph Similarity Techniques	13
2.1 Graph and Subgraph Isomorphism	14
2.2 Graph Edit Distance	17
2.3 Maximum Common Subgraph / Minimum Common Supergraph Approach	18
2.4 State Space Search Approach	21
2.5 Probabilistic Approach	22
2.6 Distance Preservation Approach	24
2.7 Relaxation Approaches	25
2.8 Mean and Median of Graphs	27
2.9 Summary	29
3. Graph Models for Web Documents	31
3.1 Pre-Processing	31
3.2 Graph Representations of Web Documents	33
3.3 Complexity Analysis	37
3.4 Web Document Data Sets	38

4. Graph-Based Clustering	41
4.1 The Graph-Based k -Means Clustering Algorithm	42
4.2 Clustering Performance Measures	44
4.3 Comparison with Previously Published Results	46
4.4 Comparison of Different Graph-Theoretical Distance Measures and Graph Representations for Graph-Based Clustering	51
4.4.1 Comparison of distance measures	52
4.4.2 Comparison of graph representations	59
4.5 Comparison of Clustering Algorithms	68
4.6 Visualization of Graph Clustering	72
4.7 The Graph-Based Global k -Means Algorithm	78
4.7.1 Global k -means vs. random initialization	80
4.7.2 Optimum number of clusters	81
5. Graph-Based Classification	87
5.1 The k -Nearest Neighbors Algorithm	88
5.1.1 Traditional method	88
5.1.2 Graph-based approach	89
5.1.3 Experimental results	89
5.2 Graph-Based Multiple Classifier Ensembles	103
5.2.1 Basic algorithm	103
5.2.2 Experimental results	105
6. The Graph Hierarchy Construction Algorithm for Web Search Clustering	109
6.1 Cluster Hierarchy Construction Algorithm (CHCA)	112
6.1.1 A review of inheritance	112
6.1.2 Brief overview of CHCA	113
6.1.3 CHCA in detail	115
6.1.4 CHCA: An example	120
6.1.5 Examination of CHCA as a clustering method	121
6.2 Application of CHCA to Search Results Processing	125
6.2.1 Asynchronous search	125
6.2.2 Implementation, input preparation and pre-processing	126
6.2.3 Selection of parameters for web search	127
6.3 Examples of Results	127

6.3.1	Comparison with Grouper	130
6.3.2	Comparison with Vivísimo	134
6.4	Graph Hierarchy Construction Algorithm (GHCA)	138
6.4.1	Parameters	138
6.4.2	Graph creation and pre-processing	140
6.4.3	Graph Hierarchy Construction Algorithm (GHCA)	141
6.4.4	GHCA examples	143
6.5	Comments	144
7.	Conclusions and Future Work	147
Appendix A	Graph Examples	151
Appendix B	List of Stop Words	215
<i>Bibliography</i>		223
<i>Index</i>		233