

Contents

Chapter I. Preliminaries	1
I.1. A brief survey of existing corpus search engines	3
Chapter II. Fundamental Data Structures	7
II.1. Notation	7
II.2. Intended use of the data structure and requirements	7
II.3. Inverted indexes	10
II.4. Tree-shaped indexes	12
II.5. Classical suffix trees	15
II.6. The distrex data structure	15
II.7. Suffix arrays	20
Chapter III. Enhancements Related to Corpus Annotation	35
III.1. Suffix arrays with annotations	35
III.2. Speed and speed improvements	54
III.3. Regular expressions and other complex patterns	56
III.4. More complex annotations	68
III.5. Results extraction and accumulators	77
III.6. Two special ways of using annotations	87
III.7. Index updates	88
III.8. Inverted files vs. suffix arrays and hybrid indexes	90
Chapter IV. Sufex On Modern Computers	95
IV.1. Parallelization	96
IV.2. Using secondary storage	102
IV.3. How to use sufex	104
Chapter V. Corpus Exploration Techniques and Applications	107
V.1. A query language for sufex	108
V.2. Harrisean methods	109
V.3. Sufex and corpus calculus	115
V.4. Building local grammars	118
Chapter VI. Results of Performance Tests	121
VI.1. Overview	121
VI.2. He-All series	124
VI.3. List22	124
VI.4. Reuters-series	124
Bibliography	129