

MST (S)
001
YANG

PAT-Tree and Local statistical information Based Chinese Phrase Extraction

By Yang YiYang

Master of Software Engineering

2009



**Faculty of Science and Technology
University of Macau**

TABLE OF CONTENTS

I.	INTRODUCTION	1
1.1.	FEATURES OF THE CHINESE DOCUMENT PROCESSING	2
1.2.	MOTIVATIONS	5
1.3.	OBJECTIVES	6
1.4.	THESIS ORGANIZATION.....	7
II.	RELATED WORKS	9
2.1.	CHINESE PHRASE EXTRACTION METHODS	9
2.2.	STATISTICS-BASED APPROACHES	11
2.3.	PAT-TREE BASED APPROACHES.....	13
III.	THE PROPOSED APPROACH.....	15
3.1.	SYSTEM PROCEDURE	15
3.2.	CONSIDER ALL POTENTIAL PHRASE PATTERNS	17
3.2.1	Semi-infinite Strings (Sistrings) and reverse Sistrings	17
3.2.2.	Drawbacks of the Sistring structure	19
3.2.3.	Longest Substring Policy	20
3.2.4.	Different Pattern Generations	22
3.3.	PAT-TREE STRUCTURE	26
3.3.1.	PAT-Tree Foundations.....	26
3.3.2.	PAT-Tree Attributes	28
3.3.3.	Attributes of PAT-Tree with All PPP	30
3.4.	LOCAL STATISTICAL INFORMATION.....	32
3.4.1.	Objective of Local statistics.....	32
3.4.2.	Local PAT-Tree Trimming	33
3.5.	PRASE EXTRACTION BASED ON GLOBAL TREE	36
IV.	EVALUATION.....	40
4.1.	RESOURCES AND PARAMETERS	42
4.2.	EXPERIMENT PARATERS SETTING	44
4.3.	METRICS	45
4.4.	INTERNAL COMPARISON	47
4.5.	EXTERNAL COMPARISON	51
V.	CONCLUSION.....	56
VI.	REFERENCES	58
	<u>Appendix</u>	60