Geographical Information Retrieval in Textual Corpora
Geographical Information Retrieval in Textual Corpora

Christian Sallaberry
Contents

**FOREWORD** .......................................................... ix
Christophe CLARAMUNT

**ACKNOWLEDGMENTS** .............................................. xi

**INTRODUCTION** .................................................. xiii

**CHAPTER 1. ACCESS BY GEOGRAPHIC CONTENT TO TEXTUAL CORPORA: WHAT ORIENTATIONS?** ........................................ 1

1.1. Introduction ..................................................... 1
1.2. Access by geographic content to textual corpora .......... 1
  1.2.1. Document retrieval and textual corpora ............... 2
  1.2.2. Textual corpora with “territorial” denotations .......... 2
  1.2.3. Access to textual content .................................. 6
1.3. Reinforcement of GIR by contributions from NLP, reasoning and multicriteria IR ...................................... 7
1.4. Toward the construction of a multicriteria IR engine .......... 9
  1.4.1. Challenges, hypotheses and research objectives .......... 10
  1.4.2. Approach .......................................................... 11
  1.4.3. Applications ...................................................... 13

**CHAPTER 2. SPATIAL AND TEMPORAL INFORMATION RETRIEVAL IN TEXTUAL CORPORA** .......................................... 17

2.1. Introduction ..................................................... 17
2.2. Review of challenges, hypotheses and research objectives ... 18
2.3. Spatial and temporal information in textual documents:
  literature review ..................................................... 19
2.3.1. Geographic information in text and IR ........................................ 19
2.3.2. Named entities ........................................................................... 19
2.3.3. Modeling languages .................................................................... 21
2.3.4. Reasoning ................................................................................... 24
2.3.5. Linguistic processing .................................................................... 26
2.3.6. GIR: systems and similarity measure models ............................... 27
2.3.7. Evaluation campaigns, corpora and resources ............................... 31
2.3.8. Summary ..................................................................................... 34

2.4. Proposition for spatial and temporal information
indexing and retrieval in textual corpora ............................................ 35
2.4.1. Reminder and focus on the notion of space and
time in “heritage” corpora ................................................................. 35
2.4.2. Core spatial model and core temporal model ............................... 36
2.4.3. Spatial and temporal relations ....................................................... 37
2.4.4. Spatial and temporal indexing process flows: PIV prototype ......... 39
2.4.5. Spatial and temporal IR: PIV prototype ....................................... 42
2.4.6. Evaluation and discussion ............................................................. 45

2.5. Summary ......................................................................................... 47
2.5.1. Contributions ............................................................................... 47
2.5.2. Perspectives ................................................................................. 49

CHAPTER 3. MULTICRITERIA INFORMATION RETRIEVAL IN
TEXTUAL CORPORA .............................................................................. 53
3.1. Introduction ..................................................................................... 53
3.2. Review of challenges, hypotheses and research objectives .............. 54
3.3. Standardization and combination of criteria: literature review ........... 56
3.3.1. Criterion standardization ............................................................... 56
3.3.2. Combination of criteria ................................................................. 58
3.3.3. Summary and positioning of a partially compensatory GIR ............ 64
3.4. Proposition for indexing by tiling and multicriteria IR in
textual corpora ..................................................................................... 65
3.4.1. Standardization by tiling ............................................................... 65
3.4.2. Spatial and temporal IR applied to tiling: PIV² ................................ 70
3.4.3. Multicriteria IR applied to tiling: PIV³ .......................................... 72
3.5. Evaluation and discussion ................................................................. 77
3.5.1. Evaluation framework of geographic IRSs: proposal
for a test collection and an experimental protocol .............................. 78
3.5.2. Evaluation of the spatial and temporal IR applied to tiling ............ 79
3.5.3. Evaluation of the multicriteria IR applied to tiling ...................... 81
3.6. Summary ......................................................................................... 84
3.6.1. Contributions ............................................................................... 84
3.6.2. Perspectives ................................................................................. 86
CHAPTER 4. GENERAL CONCLUSION .............................. 87

4.1. Summary .................................................. 87
4.1.1. Contributions to the access by geographic content to
        textual corpora ........................................... 87
4.1.2. Spatial and temporal IR in texts ....................... 88
4.1.3. Multicriteria IR in texts ............................... 89
4.2. Perspectives .............................................. 90
4.2.1. Intradimensional axis ................................ 92
4.2.2. Interdimensional axis ................................. 97
4.2.3. Expansion of the vocabulary for a qualitative
        representation of the geographic dimensions .......... 103

BIBLIOGRAPHY .................................................. 105

INDEX .......................................................... 123
This very well-documented book addresses the field of geographic information extraction and retrieval from textual documents. Geographic information retrieval from documents is, indeed, a rapidly emerging subject, a trend fostered by the growing power of the Internet and the emerging possibilities of data dissemination. Information is processed from the identification of spatial and temporal features in textual documents, data indexing and manipulation of the relevance of identified items, multicriteria retrieval and an evaluation of query results by the development of several prototypes.

The author first introduces the principles of document retrieval and then illustrates the roles and importance of spatial and temporal information in textual documents. The addressed scientific challenges lie at the intersection of information retrieval techniques, natural language processing and qualitative spatial reasoning. The contributions presented address the development of spatial and temporal data models, geographic information extraction and analysis as well as symbolic annotations. Christian Sallaberry develops several of his recent contributions oriented around the development of spatial and temporal information indexing and textual document retrieval, these propositions being, by themselves, a worthwhile contribution of this monograph.

The book is usefully completed by a rich bibliographical study of current approaches focused on the modeling and retrieval of spatial and temporal information in textual documents and similarity measures developed so far in published literature. This allows Christian Sallaberry to develop a contribution in which the linguistic annotations, as well as the developed framework, enable us to identify, interpret and retrieve spatiotemporal information. This approach is typically qualitative in the sense that the spatial and temporal features identified in a corpus can be described from spatial and temporal relationships. These relationships play an important role in the derivation of spatial and temporal indexes and the execution of information
retrieval processes, where spatial and temporal similarity measures allow us to trigger and rank query results.

The framework is completed by a multicriteria information retrieval approach. To develop and present his contribution, Christian Sallaberry introduces a useful literature review of spatiotemporal query homogenization. He introduces a spatial and temporal indexing approach based on the concepts of tiling and relevance scores, and different degrees of preference levels.

The conclusion of this book provides a broad perspective on the remaining scientific challenges. Several areas of research are discussed: integration of a domain-based ontology, modeling spatial relations in the interpretation of spatial features, generalization of these approaches in relation to the temporal and semantic dimensions, and semantic enrichment from annotations. All these domains are challenging and very attractive areas of research.

Overall, this book constitutes of a very well documented contribution, original and useful in a domain undergoing rapid development. The approach is original and brings a contribution to the field of geographic information extraction and retrieval from textual documents. It should raise wide interest for researchers in the fields of geographic and textual information processing as well as in developers of information and Web data processing systems. I hope it will generate many new vocations!

Dr Christophe CLARAMUNT
Professor in Computer Science
Naval Academy Research Institute
Lanveoc-Poullmic, June 2013
This book has its origins in my accreditation to direct research. Thus, my acknowledgments go first to Mauro Gaio, Professor at the UPPA\(^1\), for his help in this long preparation for the accreditation to direct research. I would also like to thank my reviewers Mohand Boughanem, Professor at the Paul Sabatier University in Toulouse, Christophe Claramunt, Professor at the Naval Academy of Brest, and Ross Purves, Professor at the University of Zurich, for their expert reports and numerous pieces of advice that have enabled me to improve the original manuscript. Finally, thanks go to Marie-Aude Aufaure, Professor at the Ecole Centrale de Paris, Florence Le Ber, PhD Supervisor at the Ecole nationale du génie de l’eau et de l’environnement de Strasbourg, and Thierry Nodenot, Professor at the UPPA, who have also carefully examined the manuscript of my accreditation to direct research.

I will present the results of the work conducted as a team within the laboratory of LIUPPA\(^2\). Therefore, I would like to thank once again Mauro Gaio for associating me with his research works on natural language processing and reasoning aiming at spatial and temporal information marking and analyzing in bodies of text. My thanks go equally to my colleagues at the LIUPPA. I would like to mention, in particular, my colleagues Marie-Noëlle Bessagnet, Annig Lacayrelle and Albert Royer as well as the four doctoral students Pierre Laforcade, Julien Lesbegueries, VanTien NGuyen and Damien Palacio whose works I have been able to jointly accompany: they have offered me the possibility to share in fruitful collaborations that have contributed a lot

---

2 Computer Science Laboratory of the University of Pau and Pays de l’Adour: liuppa.univ-pau.fr/.
to this research. I also address my gratitude to all other colleagues with whom I have had the chance of working within the context of different research projects. I would like to specifically thank my colleagues at the IRIT\(^3\) institute of Toulouse, Guillaume Cabanac and Gilles Hubert, for their confidence and their pertinent proposals that have contributed a lot to these results.

Without all these meetings, the work presented in this book would not have seen the light of day.

\(^3\) Institut de recherche en informatique de Toulouse: www.irit.fr/.