

# Contents

## Eighth Symposium on String Processing and Information Retrieval<sup>3/4</sup> SPIRE'2001

<b>Preface</b> _____	<b>vii</b>
<b>Program Committee</b> _____	<b>viii</b>
<b>Additional Reviewers</b> _____	<b>viii</b>
<b>Organizing Committee and Sponsors</b> _____	<b>ix</b>
<b>Regular and Invited Papers</b>	
Invited Paper: Of Maps Bigger than the Empire _____ <i>A. Apostolico</i>	2
Distributed Query Processing Using Partitioned Inverted Files _____ <i>C. Badue, R. Baeza-Yates, B. Ribeiro-Neto, and N. Ziviani</i>	10
Relating Web Characteristics with Link Based Web Page Ranking _____ <i>R. Baeza-Yates and C. Castillo</i>	21
Compaction Techniques for Nextword Indexes _____ <i>D. Bahle, H. Williams, and J. Zobel</i>	33
A Subquadratic Algorithm for Cluster and Outlier Detection in Massive Metric Data _____ <i>E. Chávez</i>	46
Speeding-up Hirschberg and Hunt-Szymanski LCS Algorithms _____ <i>M. Crochemore, C. Iliopoulos, and Y. Pinzon</i>	59
Semantic Thesaurus for Automatic Expanded Query in Information Retrieval _____ <i>M. González and V. Strube de Lima</i>	68
Exact Distribution of Deletion Sizes for Unavoidable Strings _____ <i>C. Heitsch</i>	76
On Using Two-Phase Filtering in Indexed Approximate String Matching with Application to Searching Unique Oligonucleotides _____ <i>H. Hyyrö</i>	84
On-Line Construction of Symmetric Compact Directed Acyclic Word Graphs _____ <i>S. Inenaga, H. Hoshino, A. Shinohara, M. Takeda, and S. Arikawa</i>	96
Musical Sequence Comparison for Melodic and Rhythmic Similarities _____ <i>T. Kadota, M. Hirao, A. Ishino, M. Takeda, A. Shinohara, and F. Matsuo</i>	111
A Model for the Representation and Focussed Retrieval of Structured Documents Based on Fuzzy Aggregation _____ <i>G. Kazaj, M. Lalmas, and T. Rölleke</i>	123

Evaluation of N-grams Conflation Approach in Text-Based Information Retrieval _____	136
<i>S. Kosinov</i>	
Storing Semistructured Data in Relational Databases _____	143
<i>K. Magalhães, A. Laender, and A. da Silva</i>	
Using Edit Distance in Point-Pattern Matching _____	153
<i>V. Mäkinen</i>	
Invited Paper: Re-Store: A System for Compressing, Browsing, and Searching Large Documents _____	162
<i>A. Moffat and R. Wan</i>	
Speed-up of Aho-Corasick Pattern Matching Machines by Rearranging States _____	175
<i>T. Nishimura, S. Fukamachi, and T. Shinohara</i>	
A Stemming Algorithm for the Portuguese Language _____	186
<i>V. Orenço and C. Huyck</i>	
Fast Categorisation of Large Document Collections _____	194
<i>V. Shanks and H. Williams</i>	
On Compression of Parse Trees _____	205
<i>J. Tarhio</i>	
An Efficient Bottom-Up Distance between Trees _____	212
<i>G. Valiente</i>	
Using Semantics for Paragraph Selection in Question Answering Systems _____	220
<i>J. Vicedo</i>	
Invited Paper: Semantic Labeling — Unveiling the Main Components of Meaning of Free-Text _____	228
<i>Y. Ziemann and R. Salas</i>	
<b>Poster Papers</b>	
A Comparative Study of Topic Identification on Newspaper and E-mail _____	238
<i>B. Bigi, A. Brun, J. Haton, K. Smaili, and I. Zitouni</i>	
A Documental Database Query Language _____	242
<i>N. Brisaboa, M. Penabad, A. Places, and F. Rodríguez</i>	
Design of a Graphical User Interface for Structured Documents Retrieval _____	246
<i>F. Crestani, P. de la Fuente, and J. Vegas</i>	
Genome Rearrangements Distance by Fusion, Fission, and Transposition is Easy _____	250
<i>Z. Dias and J. Meidanis</i>	
Adding Security to Compressed Information Retrieval Systems _____	254
<i>R. Milidiú, C. Gomes de Mello, and J. Fernandes</i>	
<b>Author Index _____</b>	<b>259</b>

# Preface

**S**PIRE'2001 is a Symposium on String Processing and Information Retrieval, now in its eighth edition. This series of meetings originated in South America (Belo Horizonte and Recife, Brazil, in 1993 and 1996; Valparaíso, Chile, in 1995 and 1997) and were originally called WSP (Workshop on String Processing).

Starting in 1998, at Santa Cruz, Bolivia, the focus of the workshop was broadened to include the area of information retrieval due to its increasing relevance and its inter-relationship with the area of string processing. SPIRE'1999, at Cancun, Mexico; SPIRE'2000, at A Coruña, Spain; and SPIRE'2001 continued this trend.

One of the main goals of SPIRE is to facilitate the cross fertilization between those different but related fields. I believe this objective is being met. This year we received a number of high quality contributions from the areas of string processing, information retrieval, and on problems resulting from their combination.

SPIRE'2001 was held in Laguna de San Rafael, Chile. The symposium included the presentations of 20 full papers and 5 poster papers (from 33 submitted), as well as 3 invited talks by Alberto Apostolico (Purdue University, USA, and University of Padova, Italy), Alistair Moffat (University of Melbourne, Australia) and Yuri Ziemann (Unveil Inc., USA). This year we created the category of poster papers to encourage the dissemination of promising work that still requires further development.

I wish to thank all the members of the Program Committee and the additional reviewers for their careful and mostly timely evaluation of the submissions. I would also like to thank the local organization staff for their hard and enthusiastic work, and in particular to Ricardo Baeza-Yates for compensating my inexperience as a Program Committee Chair.

**Gonzalo Navarro**  
**SPIRE'2001 Program Committee Chair**  
**Santiago, Chile**

# Program Committee

Amihood Amir, Bar-Ilan University, Israel  
Alberto Apostolico, Purdue University, USA; University of Padova, Italy  
Ricardo Baeza-Yates, Universidad de Chile, Chile  
Josep Blat, Universitat Pompeu Fabra, Spain  
Nieves Brisaboa, Universidad de A Coruña, Spain  
Edgar Chávez, Universidad Michoacana, Mexico  
Maxime Crochemore, Université de Marne-la-Vallée, France  
Russell Deaton, University of Arkansas, USA  
Pablo de la Fuente, Universidad de Valladolid, Spain  
Efthimis Efthimiadis, University of Washington, USA  
Paolo Ferragina, University of Pisa, Italy  
Edward Fox, Virginia Tech., USA  
Max Garzón, University of Memphis, USA  
John Kececioğlu, University of Arizona, USA  
João Paulo Kitajima, Universidade de Campinas, Brazil  
Eduardo Laber, Universidade Federal de Rio de Janeiro, Brazil  
Mun-Kew Leong, BIGontheNet, Singapore  
Massimo Melucci, University of Padova, Italy  
Ruy Milidiu, Pontificia Universidade Católica de Rio, Brazil  
Alistair Moffat, University of Melbourne, Australia  
Sung Hyon Myaeng, Chungnam National University, Korea  
Mario Nascimento, University of Alberta, Canada  
Gonzalo Navarro, Universidad de Chile, Chile (Chair)  
Arlindo Oliveira, INESC, Portugal  
Berthier Ribeiro-Neto, Universidade Federal de Minas Gerais, Brazil  
John Rose, University of Tokyo, Japan  
Marie-France Sagot, Institut Pasteur, France  
João Setubal, Universidade de Campinas, Brazil  
Masayuki Takeda, Kyushu University, Japan  
Jorma Tarhio, Helsinki University of Technology, Finland  
Esko Ukkonen, University of Helsinki, Finland  
Nivio Ziviani, Universidade Federal de Minas Gerais, Brazil

## Additional Reviewers

Charles Ornelas Almeida  
Claudine Santos Badue  
Alejandro Bassi  
Daniela Ferreira da Mota  
Alejandro Hevia  
Costas S. Iliopoulos  
Juha Kärkkäinen  
Pekka Kilpeläinen  
Shmuel T. Klein  
Stuart Margolis  
Edleno Silva de Moura  
Fernando das Neves  
Artur Alves Pessoa  
Mathieu Raffinot  
Davood Rafiei  
Raul Renteria  
Altigran Soares da Silva  
Jesús Vegas

# Organizing Committee

Ricardo Baeza-Yates, Univ. de Chile, Chile  
Edgar Chávez, Univ. Michoacana, Mexico  
Gonzalo Navarro, Univ. de Chile, Chile (Chair)  
Cuauhtemoc Rivera-Loaiza, Univ. Michoacana, Mexico

## Sponsors

CYTED Project VII.19 RIBIDI (Recuperación de Información y Bibliotecas Digitales)  
SCCC (Sociedad Chilena de Ciencia de la Computación)  
Universidad de Chile, Santiago, Chile  
Universidad Michoacana de San Nicolas de Hidalgo, Morelia, Mexico