Preface

The papers contained in this volume were presented at the twelfth edition of the International Symposium on String Processing and Information Retrieval (SPIRE), held November 2–4, 2005 in Buenos Aires, Argentina. They were selected from 102 papers submitted from 25 countries in response to the call for papers. A total of 27 submissions were accepted as full papers, yielding an acceptance rate of about 26%. In view of the large number of good-quality submissions the conference program also included 17 short papers that also appear in the Proceedings. In addition, the Steering Committee invited the following speakers: Prabhakar Raghavan (Yahoo! Research, USA), Paolo Ferragina (University of Pisa, Italy), and Gonzalo Navarro (University of Chile, Chile).

Papers solicited for SPIRE 2005 were meant to constitute original contributions to areas such as string processing (dictionary algorithms, text searching, pattern matching, text compression, text mining, natural language processing, and automata based string processing); information retrieval languages, applications, and evaluation (IR modeling, indexing, ranking and filtering, interface design, visualization, cross-lingual IR systems, multimedia IR, digital libraries, collaborative retrieval, Web related applications, XML, information retrieval from semi-structured data, text mining, and generation of structured data from text); and interaction of biology and computation (sequencing and applications in molecular biology, evolution and phylogenetics, recognition of genes and regulatory elements, and sequence driven protein structure prediction).

SPIRE has its origins in the South American Workshop on String Processing (WSP). Since 1998 the focus of the conference was broadened to include information retrieval. Starting in 2000, Europe has been the conference venue on even years. The first 11 meetings were held in Belo Horizonte (Brazil, 1993), Valparaíso (Chile, 1995), Recife (Brazil, 1996), Valparaíso (Chile, 1997), Santa Cruz (Bolivia, 1998), Cancún (Mexico, 1999), A Coruña (Spain, 2000), Laguna San Rafael (Chile, 2001), Lisboa (Portugal, 2002), Manaus (Brazil, 2003), and Padova (Italy, 2004).

SPIRE 2005 was held in tandem with LA-WEB 2005, the Third Latin American Web Congress, with both conferences sharing a common day in Web Retrieval.

SPIRE 2005 was sponsored by Centro Latinoamericano de Estudios en Informática (CLEI), Programa Iberoamericano de Ciencia y Tecnología para el Desarrollo (CYTED), Center for Web Research (CWR, University of Chile), and Sociedad Argentina de Informática e Investigación Operativa (SADIO).

We thank the local organizers for their support in the organization of SPIRE and the members of the Program Committee and the additional reviewers for providing timely and detailed reviews of the submitted papers and for their active participation in the email discussions that took place before we could assemble the final program. Finally, we would like to thank Ricardo Baeza-Yates, who, on behalf of the Steering Committee, invited us to chair the Program Committee.

Mariano P. Consens
Gonzalo Navarro
SPIRE 2005 Program Chairs
Buenos Aires, November 2005

SPIRE 2005 Organization

Steering Committee

Ricardo Baeza-Yates (Chair)	ICREA-Universitat Pompeu Fabra (Spain) and Universidad de Chile (Chile)
Alberto Apostolico	Università di Padova (Italy) and Georgia Tech (USA)
Alberto Laender	Universidade Federal de Minas Gerais (Brazil)
Massimo Melucci	Università di Padova (Italy)
Edleno de Moura	Universidade Federal do Amazonas (Brazil)
Mario Nascimento	University of Alberta (Canada)
Arlindo Oliveira	INESC (Portugal)
Berthier Ribeiro-Neto	Universidade Federal de Minas Gerais (Brazil)
Nivio Ziviani	Universidade Federal de Minas Gerais (Brazil)

Program Committee Chairs

Mariano Consens	Dept. of Mechanical and Industrial Engineering Dept. of Computer Science University of Toronto, Canada
Gonzalo Navarro	Center for Web Research Dept. of Computer Science Universidad de Chile, Chile

Program Committee Members

Amihood Amir	Bar-Ilan University (Israel)
Alberto Apostolico	Università di Padova (Italy) and Georgia Tech (USA)
Ricardo Baeza-Yates	ICREA-Universitat Pompeu Fabra (Spain) and Universidad de Chile (Chile)
Nieves R. Brisaboa	Universidade da Coruña (Spain)
Edgar Chávez	Universidad Michoacana (Mexico)
Charles Clarke	University of Waterloo (Canada)
Bruce Croft	University of Massachussetts (USA)
Paolo Ferragina	Università di Pisa (Italy)
Norbert Fuhr	Universität Duisburg-Essen (Germany)
Raffaele Giancarlo	Università di Palermo (Italy)
Roberto Grossi	Università di Pisa (Italy)
Carlos Heuser	Universidade Federal de Rio Grande do Sul (Brazil)
Carlos Hurtado	Universidad de Chile (Chile)
Lucian Ilie	University of Western Ontario (Canada)
Panagiotis Ipeirotis	New York University (USA)
Juha Kärkkäinen	University of Helsinki (Finland)
Nick Koudas	University of Toronto (Canada)
Mounia Lalmas	Queen Mary University of London (UK)
Gad Landau	University of Haifa (Israel) and Polytechnic University (NY, USA)
Stefano Lonardi	University of California (USA)
Yoelle Maarek	IBM Haifa Research Lab (Israel)
Veli Mäkinen	Bielefeld University (Germany)
Mauricio Marín	Universidad de Magallanes (Chile)
João Meidanis	UNICAMP (Brazil)
Massimo Melucci	Università di Padova (Italy)
Edleno de Moura	Universidade Federal do Amazonas (Brazil)
Ian Munro	University of Waterloo (Canada)
Arlindo Oliveira	INESC (Portugal)
Kunsoo Park	Seoul National University (Korea)
Prabhakar Raghavan	Yahoo Inc. (USA)
Berthier Ribeiro-Neto	Universidade Federal de Minas Gerais (Brazil)
Kunihiko Sadakane	Kyushu University (Japan)
Marie-France Sagot	INRIA (France)
João Setubal	Virginia Tech (USA)
Jayavel Shanmugasundaram	Cornell University (USA)
Ayumi Shinohara	Tohoku University (Japan)
Jorma Tarhio	Helsinki University of Technology (Finland)
Jeffrey Vitter	Purdue University (USA)
Hugh Williams	Microsoft Corporation (USA)
Hugo Zaragoza	Microsoft Research (UK)
Nivio Ziviani	Universidade Federal de Minas Gerais (Brazil)
Justin Zobel	RMIT (Australia)

External Reviewers

Jussara Almeida
Ramurti Barbosa
Sebastian Böcker
David Carmel
Marco Cristo
Alair Pereira do Lago
Celia Francisca dos Santos
Feng Shao
Kimmo Fredriksson
Paulo B. Golgher
Jie Zheng
Shahar Keret
Sascha Kriewel
Nicholas Lester
Julia Mixtacki
Henrik Nottelmann
Rodrigo Paredes
Hannu Peltola
Nadia Pisanti
Bruno Possas
Davi de Castro Reis
Luis Russo
Marinella Sciortino
Darren Shakib
S.M.M. (Saied) Tahaghoghi
Andrew Turpin
Ying Zhang
Michela Bacchin
Bodo Billerbeck
Michael Cameron
Luis Coelho
Giorgio Maria Di Nunzio
Shiri Dori
Fan Yang
Nicola Ferro
Gudrun Fisher
Alejandro Hevia
Carmel Kent
Tsvi Kopelowitz
Michael Laszlo
Saadia Malik
Viviane Moreira Orengo
Nicola Orio
Laxmi Parida
Patrícia Peres
Benjamin Piwowarski
Jussi Rautio
Nora Reyes
Klaus-Bernd Schürmann
Rahul Shah
Riva Shalom
Eric Tannier
Rodrigo Verschae

Local Organization

SADIO (Argentine Society for Informatics and Operations Research)

SADIO President:	Gabriel Baum
Local Arrangements Chair:	Héctor Monteverde
Steering Committee Liaison:	Ricardo Baeza-Yates
Administrative Manager:	Alejandra Villa

Table of Contents

String Processing and Information Retrieval 2005

Enhanced Byte Codes with Restricted Prefix Properties 1
J. Shane Culpepper, Alistair Moffat (University of Melbourne, Australia)

Experimental Analysis of a Fast Intersection Algorithm for Sorted Sequences 13
Ricardo Baeza-Yates, Alejandro Salinger (University of Chile, Chile)

Compressed Perfect Embedded Skip Lists for Quick Inverted-Index Lookups 25
Paolo Boldi, Sebastiano Vigna (Università degli Studi di Milano, Italy)

XML Retrieval with a Natural Language Interface 29
Xavier Tannier (Ecole Nationale Supérieure des Mines de Saint-Etienne, France), Shlomo Geva (Queensland University of Technology, Australia)

Recommending Better Queries from Click-Through Data 41
Georges Dupret (Universidad de Chile, Chile), Marcelo Mendoza (Universidad de Valparaíso, Chile)

A Bilingual Linking Service for the Web 45
Alessandra Alaniz Macedo (Sao Paulo University, Brazil), José Antonio Camacho-Guerrero (3WT, Brazil), Maria da Graça Campos Pimentel (Sao Paulo University, Brazil)

Evaluating Hierarchical Clustering of Search Results 49
Juan M. Cigarran, Anselmo Peñas, Julio Gonzalo, Felisa Verdejo (UNED, Spain)

Counting Suffix Arrays and Strings 55
Klaus-Bernd Schürmann, Jens Stoye (Universität Bielefeld, Germany)

Towards Real-Time Suffix Tree Construction 67
Amihood Amir (Bar-Ilan University, Israel; Georgia Tech, USA), Tsvi Kopelowitz, Moshe Lewenstein (Bar-Ilan University, Israel), Noa Lewenstein (Netanya College, Israel)

Rank-Sensitive Data Structures 79
Iwona Bialynicka-Birula, Roberto Grossi (Università di Pisa, Italy)

Cache-conscious Collision Resolution in String Hash Tables 91
Nikolas Askitis, Justin Zobel (RMIT University, Australia)

Measuring the Difficulty of Distance-Based Indexing 103
Matthew Skala (University of Waterloo, Canada)

N-gram Similarity and Distance 115
Grzegorz Kondrak (University of Alberta, Canada)

Using the k-Nearest Neighbor Graph for Proximity Searching in Metric Spaces 127
Rodrigo Paredes (University of Chile, Chile), Edgar Chávez (Universidad Michoacana, Mexico)

Classifying Sentences using Induced Structure 139
Menno van Zaanen, Luiz Augusto Pizzato, Diego Mollá (Macquarie University, Australia)

Counting Lumps in Word Space: Density as a Measure of Corpus Homogeneity 151
Magnus Sahlgren, Jussi Karlgren (Swedish Institute of Computer Science, Sweden)

Multi-label Text Categorization Using K-Nearest Neighbor Approach with M-Similarity 155
Yi Feng, Zhaohui Wu, Zhongmei Zhou (Zhejiang University, China)

Lydia: A System for Large-Scale News Analysis 161
Levon Lloyd, Dimitrios Kechagias, Steven Skiena (State University of New York at Stony Brook, USA)

Composite Pattern Discovery for PCR Application 167
Stanislav Angelov (University of Pennsylvania, USA), Shunsuke Inenaga (Kyushu University, Japan)

Lossless Filter for Finding Long Multiple Approximate Repetitions Using a New Data Structure, the Bi-Factor Array 179
Pierre Peterlongo (Université de Marne-la-Vallée, France), Nadia Pisanti (Università di Pisa, Italy; Université Paris-Nord, France), Frederic Boyer (INRIA Rhône-Alpes and Université Claude Bernard, France), Marie-France Sagot (INRIA Rhône-Alpes and Université Claude Bernard, France; King's College London, UK)

Linear Time Algorithm for the Generalised Longest Common Repeat Problem 191
Inbok Lee, Yoan José Pinzón Ardila (King's College London, UK)

Application of Clustering Technique in Multiple Sequence Alignment 202
Patrícia Silva Peres, Edleno Silva de Moura (Universidade Federal do Amazonas, Brazil)

Stemming Arabic Conjunctions and Prepositions 206
Abdusalam F.A. Nwesri, S.M.M. Tahaghoghi, Falk Scholer (RMIT University, Australia)

XML Multimedia Retrieval 218
Zhigang Kong, Mounia Lalmas (Queen Mary University of London, UK)

Retrieval Status Values in Information Retrieval Evaluation 224
Amélie Imafouo, Xavier Tannier (Ecole Nationale Supérieure des Mines de Saint-Etienne, France)

A Generalization of the Method for Evaluation of Stemming Algorithms Based on Error Counting 228
Ricardo Sánchez de Madariaga, José Raúl Fernández del Castillo, José Ramón Hilera (University of Alcalá, Spain)

Necklace Swap Problem for Rhythmic Similarity Measures 234
Yoan José Pinzón Ardila (King's College London, UK), Raphaël Clifford (University of Bristol, UK), Manal Mohamed (King's College London, UK)

Faster Generation of Super Condensed Neighbourhoods using Finite Automata 246
Luís M. S. Russo, Arlindo L. Oliveira (IST/INESC-ID, Portugal)

Restricted Transposition Invariant Approximate String Matching Under Edit Distance 257
Heikki Hyyrö (University of Tampere, Finland)

Fast Plagiarism Detection System 268
Maxim Mozgovoy, Kimmo Fredriksson (University of Joensuu, Finland), Daniel White, Mike Joy (University of Warwick, UK), Erkki Sutinen (University of Joensuu, Finland)

A Model for Information Retrieval based on Possibilistic Networks 272
Asma H. Brini, Mohand Boughanem, Didier Dubois (IRIT, France)

Comparison of Representations of Multiple Evidence using a Functional Framework for IR 284
Ilmério R. Silva, João N. Souza, Luciene C. Oliveira (Federal University of Uberlândia, Brazil)

Deriving TF-IDF as a Fisher kernel 296
Charles Elkan (University of California at San Diego, USA)

Utilizing Dynamically Updated Estimates in Solving the Longest Common Subsequence Problem 302
Lasse Bergroth (TUCS and Turku University, Finland)

Computing Similarity of Run-Length Encoded Strings with Affine Gap Penalty 314
Jin Wook Kim (Seoul National University, Korea), Amihood Amir (Bar-Ilan University, Israel; Georgia Tech, USA), Gad M. Landau (University of Haifa, Israel; Polytechnic University of New York, USA), Kunsoo Park (Seoul National University, Korea)

L_1 Pattern Matching Lower Bound 326
Ohad Lipsky, Ely Porat (Bar-Ilan University, Israel)

Approximate Matching in the L_∞ Metric 330
Ohad Lipsky, Ely Porat (Bar-Ilan University, Israel)

An Edit Distance between RNA Stem-loops 334
Valentin Guignon (Université de Montréal, Canada), Cedric Chauve (Université du Québec à Montréal, Canada), Sylvie Hamel (Université de Montréal, Canada)

A Multiple Graph Layers Model with Application to RNA Secondary Structures Comparison 346
Julien Allali (Université de Marne-la-Vallée, France), Marie-France Sagot (INRIA Rhône-Alpes and Université Claude Bernard, France; King's College London, UK)

Normalized Similarity of RNA Sequences 358
Rolf Backofen (Friedrich-Schiller Universität Jena, Germany), Danny Hermelin (University of Haifa, Israel), Gad M. Landau (University of Haifa, Israel; Polytechnic University of New York, USA), Oren Weimann (University of Haifa, Israel)

A Fast Algorithmic Technique for Comparing Large Phylogenetic Trees . . 368
Gabriel Valiente (Technical University of Catalonia, Spain)

Practical and Optimal String Matching 374
Kimmo Fredriksson (University of Joensuu, Finland), Szymon Grabowski (Technical University of Łódź, Poland)

A Bit-parallel Tree Matching Algorithm for Patterns with Horizontal VLDC's 386
Hisashi Tsuji, Akira Ishino (Kyushu University, Japan), Masayuki Takeda (Kyushu University and JST, Japan)

A Partition-Based Efficient Algorithm for Large Scale Multiple-Strings Matching 397
Liu Ping, Liu Yan-bing, Tan Jian-long (Chinese Academy of Sciences, China)

Author Index 403