IADIS INTERNATIONAL CONFERENCE

applied computing 2008

10 - 13 April
Algarve, Portugal

Proceedings of IADIS International Conference Applied Computing 2008

Edited by
Nuno Guimarães
Pedro Isaías

iadis
international association for development of the information society

IADIS INTERNATIONAL CONFERENCE

APPLIED COMPUTING 2008

PROCEEDINGS OF THE
IADIS INTERNATIONAL CONFERENCE
APPLIED COMPUTING 2008

ALGARVE, PORTUGAL

10-13 April 2008

Organised by
IADIS
International Association for Development of the Information Society

Edited by Nuno Guimarães and Pedro Isaías

Associate Editors: Luís Rodrigues and Patrícia Barbosa

ISBN: 978-972-8924-56-0

SUPPORTED BY

FCT Fundação para a Ciência e a Tecnologia

MINISTÉRIO DA CIÊNCIA E DO ENSINO SUPERIOR Portugal

TABLE OF CONTENTS

FULL PAPERS

SHORT PAPERS

POSTERS

FOREWORD

These proceedings contain the papers of the IADIS Applied Computing 2008, which was organised by the International Association for Development of the Information Society in Algarve, Portugal, 10-13 April 2008.

The IADIS Applied Computing conference aims to address the main issues of concern within the applied computing area and related fields. This conference covers essentially technical aspects.

The following thirty-three areas have been object of paper and poster submissions:

Agent Systems and Applications; Algorithms; Applied Information Systems; Case Studies and Applications; Communications; Data Mining; Database Systems; E-Commerce Theory and Practice; Embedded Systems; Evaluation and Assessment; Global Tendencies; Information Retrieval; Intelligent Systems; Mobile Networks and Systems; Multimedia; Networking; Object Orientation; Parallel and Distributed Systems; Payment Systems; Programming Languages; Protocols and Standards; Semantic Web; Software Engineering; Storage Issues; Technologies for E-Learning; Wireless Applications; WWW Applications; WWW Technologies; Ubiquitous Computing; Usability Issues; Virtual Reality; Visualization; XML and other Extensible Languages.

The IADIS Applied Computing 2008 Conference had 217 submissions from 37 countries. Each submission has been anonymously reviewed by an average of 4 independent reviewers, to ensure the final high standard of the accepted submissions. Out of the papers submitted, 39 got blind referee ratings that published them as full papers, which means that the acceptance rate was below 20 %. Some other submissions were published as reflection papers, short papers and posters. The best papers will be selected for publishing as extended versions in the IADIS Journal on Computer Science and Information Systems and other selected journals.

The conference, besides the presentation of full papers, reflection papers, short papers and posters also includes one keynote presentation from an internationally distinguished researcher: we wish to thank Dr. Marcin Paprzycki, Systems Research Institute Polish Academy of Science, Poland. Also a special thanks to the conference tutorial given by Dr. Marcin Paprzycki and Dr. Maria Ganzha, Systems Research Institute Polish Academy of Science, Poland.

As we all know, a conference requires the effort of many individuals. We would like to thank all members of the Program Committee (163 top researchers in their fields) for they hard work in reviewing and selecting the papers that appear in the book of the proceedings. Special thanks also to the auxiliary reviewers that contributed to the reviewing process.

Last but not the least, we hope that everybody will have a good time in Algarve, and we invite all participants for the next year edition of the IADIS International Conference Applied Computing 2009.

Nuno Guimarães, Faculdade de Ciências - University of Lisbon, Portugal
Pedro Isaías, Universidade Aberta (Portuguese Open University), Portugal
Conference and Program Co-Chairs

Algarve, Portugal
10 April 2008

PROGRAM COMMITTEE

CONFERENCE AND PROGRAM CO-CHAIRS

Nuno Guimarães, Faculdade de Ciências - University of Lisbon, Portugal
Pedro Isaías, Universidade Aberta (Portuguese Open University), Portugal

COMMITTEE MEMBERS

Abdelhamid Mellouk, University of Paris XII, France
Abdelmajid Kadri, ENSAM, France
Achim Basermann, NEC Europe Ltd., Germany
Adam Adamopoulos, Democritus University of Thrace, Greece
Adam Wong, Deakin University, Australia
Adam Wojciechowski, Poznan University of Technology, Poland
Aijuan Dong, Hood College, USA
Alexander Lazovik, University of Trento, Italy
Ali Asghar Shiri, University of Alberta, Canada
Ali Masoudi-Nejad University of Tehran, Iran
Anders Gidenstam, Max Planck Inst. For Comp. Science, Germany
Andreas Andreou, University of Cyprus, Cyprus
Andreas Wombacher, University of Twente, The Netherlands
Ana María Sánchez Montero. Universidad Carlos III de Madrid, Spain
Anna Maddalena, Università degli Studi di Genova, Italy
Annamaria Chiasera, University of Trento, Italy
Antonino Sabetta, ISTI-CNR, Italy
Antonio Bucchiarone, IMT of Lucca, Italy
Anton Michlmayr, Vienna University of Technology, Austria
Arndt Bode, Technical University München, Germany
Arno Wacker, University of Duisburg-Essen, Germany
Aurelio Bermúdez, Universidad de Castilla-La Mancha, Spain
Azzedine Benameur, SAP Research, France
Baoying Wang, Waynesburg University, USA
Beda Christoph Hammerschmi, Institute of Information Systems, Germany
Bin Luo, Anhui University, China
Blanca Caminero, Universidad de Castilla-La Mancha, Spain
Boris Epstein, The Academic College of Tel-Aviv-Yaffo, Israel
Boris Koldehofe, University of Stuttgart, Germany
Brian Kirkegaard, Royal School of Library & Information Science, Denmark
Carmen Ruiz, Universidad de Castilla-La Mancha, Spain
Chao Peng, Japan Advanced Institute of Science and Technology (JAIST), Japan
Ching-Hsien Hsu, Chung Hua University, Taiwan
Daniel S. Katz, Louisiana State University, USA

Daniele Radicioni, Turin University, Italy
Daoqiang Zhang, Nanjing University of Aeronautics and Astronautics, China
Dick Stenmark, Göteborg University, Sweden
Dimitris Kalles, Hellenic Open University, Greece
Dirk Koschützki, IPK, Germany
Djamila Ouelhadj, University of Nottingham, United Kingdom
Dongmei Ren, IBM Silicon Valley Lab, USA
Dragoljub Pokrajac, Delaware State University, USA
Eda Marchetti, ISTI-CNR, Italy
Efstratios Georgopoulos, Technical University of Kalamata, Greece
Elarbi Badidi, United Arab Emirates University, United Arab Emirates
Elie El Khoury, University of Groningen, The Netherlands
Emmanuel Karapidakis, Department of Electronics / T.E.I. of Crete, Greece
Enrique Árias, Universidad de Castilla-La Mancha, Spain
Falk Schreiber, IPK, Germany
Farid Naït-Abdesselam, University of Lille, France
Federico Bergenti, University of Parma, Italy
Filippos Azariadis, University of the Aegean, Greece
Florian Rosenberg, Vienna University of Technology, Austria
Fotis Liarokapis, Coventry University, UK
Francesca Lonetti, ISTI-CNR, Italy
Francesca Martelli, University of Pisa, Italy
Francisco Garcia Peñalvo, University of Salamanca, Spain
Francisco Parreño Torres, Universidad de Castilla-La Mancha, Spain
Ganna Frankova, University of Trento, Italy
Gareth Jones, Dublin City University, Ireland
Georgina Gaughan, School of Computing, Dublin City University, Ireland
Giacomo Cabri, Universita' di Modena e Reggio Emilia, Italy
Gilles Hubert, Université Paul Sabatier, France
Gongzhu Hu, Central Michigan University, USA
Gregor Schiele, University of Mannheim, Germany
Grigorios Beligiannis, University of Ioannina, Greece
Guadalupe Ortiz Bellot, University of Extremadura, Spain
Gugliemo De Angelis, Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo", Italy
Hai Jin, Huazhong University of Science and Technology, P. R. China
Hamid Arabnia, University of Georgia, USA
Hind Castel, Institut National des Télécommunications, France
Holger D. Hofmann, ekaabo, Germany
Hongwu Ma, Edingburgh University, Germany
Imad Rahal, College of Saint Benedict | Saint John's University, USA
Ina Koch, Technical University of Applied Sciences Berlin, Germany
Ivan Jelinek, Czech Technical University, Czech Republic
Jalel Benothman, University of Versailles, France
Jason Hung, Overseas Chinese Institute of Technology, Taiwan
Jerry Hsi-Ya Chang, Nat'l Center for High Performance Computing, Taiwan
Jiann-Liang Chen, National Dong Hwa University, Taiwan
Jie Hu, St Cloud State University, USA
Jie Tao, Universität Karlsruhe, Germany

Jinghua Gao, ISTI-CNR, Italy
Jixin Ma, Greenwich University, UK
Jo Abrantes, University of Wollongong, Australia
Johannes Meinecke, University of Karlsruhe, Germany
Jörg Hähner, University of Hannover, Germany
José Manuel Molina López, Universidad Carlos III de Madrid, Spain
José M Peña, Technical University of Madrid, Spain
Juan J. Pardo, Universidad de Castilla-la-Mancha, Spain
Juan José Sánchez Peña, Universidad de Alcalá, Spain
Juan Manuel Fernández Luna, University of Granada, Spain
Julian Padget, University of Bath, UK
Julio Calvo. Universidad de Alcalá, Spain
Ken Barker, University of Calgary, Canada
Koen Bertels, TU Delft, The Netherlands
Kuan-Ching Li, Providence University, Taiwan
Lynda Mokdad, Université Paris Dauphine, France
Manuel E. Acacio Sánchez, Universidad de Murcia, Spain
Marco De Gemmis, University of Bari, Italy
Marcus Handte, University of Bonn, Germany
Maria Damiani, Università degli Studi di Milano, Italy
María N. Moreno García, Universidad de Salamanca, Spain
Mario Donato Marino, Universidade de Sao Paulo, Brazil
Martin Fredriksson, Blekinge Institute of Technology, Sweden
Martin Knahl, University of Plymouth, United Kingdom
Matthias Lange, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany
Max Chevalier, Université de Toulouse, France
Mei-Ling Shyu, University of Miami, USA
Mercedes G. Merayo, Universidad Complutense de Madrid, Spain
Miguel Angel Patricio Guisado, Universidad de Alcalá, Spain
Min-Ling Zhang, HoHai University, China
Miroslav Bures, Czech Technical University, Czech Republic
Miroslaw Staron, IT University of Göteborg, Sweden
Nashat Mansour, Lebanese American University, Lebanon
Natallia Kokash, University of Trento, Italy
Nataliya Rassadko, Università degli Studi di Trento, Italy
Necip Sahinkaya, University of Bath, UK
Nikola Stojanovic, University of Texas at Arlington, USA
Nikolaos Sapidis, University of the Aegean, Greece
Olivier Teste, Université Paul Sabatier, France
Ouri Wolfson, University of Illinois at Chicago, USA
Pascal Lorenz, University of Haute Alsace, France
Paul EL Khoury, SAP Research, France
Pedro García, Universidad de Castilla-La Mancha, Spain
Pierre Busnel, Sherbrooke University, Canada
Phuong Hoai Ha, University of Tromsø, Norway
Qin Ding, East Carolina University, USA
Quan Thanh Tho, Hochiminh City University of Technology, Vietnam
Rafa Al-Qutaish, Applied Science University in Amman, Jordan
Rafael Casado, University of Castilla-La Mancha, Spain

Rami Yared, Japan Advanced Institute of Science and Technology (JAIST), Japan
Raúl Suárez, Universitat Politècnica de Catalunya, Spain
Richard Lai, La Trobe University, Australia
Robert Wrembel, Poznan University Of Technology, Poland
Rodrigo Fernandes de Mello, Universidade de Sao Paulo, Brazil
Ruy Jauregui, BioBase, Germany
Sergey Peigin, Israel Aircraft Industries, Israel
Sharon Cox, University of Central England, United Kingdom
Shu-Ching Chen, Florida International University, USA
Simon Richir, ENSAM, France
Sotirios Terzis, University of Strathclyde, United Kingdom
Spyros Vosinakis, University of the Aegean, Greece
Stefanos Mavromoustakos, School of Computer Science and Engineering Cyprus College, Cyprus
Steffen Neumann, Leibniz Institute of Plant Biochemistry, Germany
Stéphane Maag, National Institute of Telecommunications, France
Susumu Goto, Kyoto University, Japan
Tarek Bejaoui, Mediatron, Sup'Com, Tunisia
Ting-Wei Hou, National Cheng-Kung University, Taiwan
Tony Gorschek, Blekinge Institute of Technology, Sweden
Urszula Markowska-Kaczmar, Wroclaw University of Technology, Poland
Vassilis Delis, Research Academic Computer Technology Institute, Greece
Victor Robles, Technical University of Madrid, Spain
Vincenzo Deufemia, Università di Salerno, Italy
Wenbin Jiang, Huazhong University of Science and Technology, China
Wenbing Tao, Huazhong University of Science and Technology, China
Xiang Fu, Georgia Southwestern State University, USA
Xiao-Lin Li, Nanjing University, China
Xiaoyang Tan, Nanjing University of Aeronautics and Astronautics, China
Xinlian Liu, Assistant Professor of Hood College, USA
Yoshifumi Manabe, NTT Cyber Space Laboratories, Japan
Zhi-Hua Zhou, Nanjing University, China

AUXILIARY REVIEWERS

David Levine, University of Texas at Arlington, USA
Gregorio Diaz Descalzo, Universidad de Castilla-La Mancha, Spain
Mario Giacobini, Turin University, Italy
Matthew Wright, University of Texas at Arlington, USA
Olga Zlydareva, University of Trento, Italy

TEXT-MINING RESEARCH IN GENOMICS

Carmen Galvez
University of Granada
Communication and Documentation Faculty, Granada 18071, Spain

Félix Moya-Anegón
University of Granada / Spanish National Research Council (CSIC)
Communication and Documentation Faculty, Granada 18071, Spain

ABSTRACT

Biomedical text-mining have great promise to improve the usefulness of genomic researchers. The goal of text-mining is analyzed large collections of unstructured documents for the purposes of extracting interesting and non-trivial patterns of knowledge. The analysis of biomedical texts and available databases, such as Medline and PubMed, can help to interpret a phenomenon, to detect gene relations, or to establish comparisons among similar genes in different specific databases. All these processes are crucial for making sense of the immense quantity of genomic information. In genomics, text-mining research refers basically to the creation of literature networks of related biological entities. Text data represent the genomics knowledge base and can be mined for relationships, literature networks, and new discoveries by literature relational chaining. However, text-mining is an emerging field without a clear definition in the genomics. This work presents some applications of text-mining to genome-based research, such as the genomic term identification in curation processes, the formulation of hypotheses about disease, the visualization of biological relationships, or the life-science domain mapping.

KEYWORDS

Text-mining; Genomics; Bioinformatics; Knowledge Discovery in Text (KDT)

1. INTRODUCTION

The volume of published biomedical research, and therefore electronically available databases, is growing at an unprecedented rate, making it hard for life-science researchers to stay up-to-date. Due to the overload of information, biomedical scientists are faced with major challenges when tracking down new discoveries and the results of research in their domain of interest. These challenges are intensified by the need to follow developments in other domains that might possibly be relevant to one's own research. Comparative genomics take in epidemiology, clinical diagnosis, the development of new drugs and of DNA-based genetic tests. When researchers cannot build on each other's experiments, scientific progress may be slowed or research may be needlessly duplicated.

The most of what is known about genes and genomes is to be uncovered in the biomedical literature (Yandell & Majoros, 2002). Current expansion has heightened interest in: (a) *Information Retrieval* (IR), to gather, select, and filter documents that may prove useful; (b) *Natural Language Processing* (NLP) to automatically process the texts; and (c) *Information Extraction* (IE), a sub-area of NLP, to find relevant concepts, facts surrounding concepts, and relationships between relevant terms from the identified documents. There has been a lot of activity in the field of text-mining in biology. The Text REtrieval Conference (TREC) implemented a Genomic Track to create an experimental environment for research in the use of information retrieval systems in the genomic domain (Hersh, 2005). NLP has also attracted attention at bioinformatics meetings in recent years, such as the *Intelligent Systems for Molecular Biology* (ISMB), *European Conference on Computational Biology* (ECCB) and the *Pacific Symposium on Biocomputing* (PSB). *BioCreAtIvE* (Critical Assessment of Information Extraction in Biology) is a forum to discuss results of NLP tasks applied to biomedical literature.

2. TEXT-MINING APPROACHES

Text-mining has its origin from data-mining. The information in conventional data-mining is usually highly structured, containing mostly numbers and symbols. Data-mining is an analytical process entailing IR, NLP and IE, used to discover unsuspected associations – that is, combining or linking facts and events for the purpose of *Knowledge Discovery in Databases* (KDD). Data-mining methods can be generally grouped as: (a) *supervised methods*, to present documents according to predefined classes, such as techniques for inserting new documents into a previously existing ontology; and (b) *unsupervised methods*, such as clustering algorithms and visualization techniques, which gather texts on the basis of their similarity and thereby reduce the dimensionality of text representation.

When data-mining processes are applied to texts in natural language, we speak of text-mining, also known as textual data-mining, intelligent text analysis, text data-mining, unstructured data management, or *Knowledge Discovery in Text* (KDT). Text-mining, then, is the discovery by computer of previously unknown information, through the automatic extraction of information from different written resources. A key element is the linking together of this extracted information to form new facts or hypotheses that can be further explored using more conventional experimental means. Text-mining is the process of discovering and extracting knowledge from unstructured data, contrasting it with data-mining which discovers knowledge from structured data (Hearst, 1999). Text-mining enables analysis of large collections of unstructured documents for the purposes of extracting interesting and non-trivial patterns of knowledge.

Biomedical knowledge in literature can be discovered through three basic procedures (Leroy & Chen, 2005): (*i*) top-down approaches, where researchers form hypotheses that lead to specific experiments, or create ontologies to describe the terminology and knowledge common to a given domain; (*ii*) bottom-up approaches, which try to discover interesting patterns or associations in existing data, in turn used to form new hypotheses (clustering techniques are used frequently for this purpose); and (*iii*) hybrid methods, involving several techniques and knowledge sources in combination, such as information retrieval and term co-occurrence analysis, to arrive at complementary sets of documents that can help researchers articulate new hypotheses. In many cases implicit relationships are inferred simply by combining the principle of the co-occurrence of terms or concepts to some form of graphic association. Biomedical text-mining is organized in stages classified into the following five steps:

Step 1. *Text gathering*: the process of text gathering in biomedical literature is largely dominated by Medline[1] and PubMed[2].

Step 2. *Text pre-processing*: biomedical texts are analyzed and stored in an internal representation form, after the elimination of stop-words, the exclusion of overly frequent terms, term standardization via stemming or lemmatisation, and the detection of noun phrases. Also text processing means tokenisation and then part-of-speech tagging, entity tagging or labelling and term recognition. Biomedical text pre-processing means tokenization and biological entities tagging, or in a *bag-of-words* approach word stemming. Biomedical text-mining uses techniques from the field of data-mining but, because it deals with unstructured data, a major part of the text-mining process revolves around the crucial stage of pre-processing the document collections. NLP plays a major role in text-mining as it transforms text into structures that can be analyzed statistically.

Step 3. *Text analysis or text categorization (clustering or classification)*: textual data can be analysed using text-mining algorithms, that is, applying either unsupervised or supervised methods. Data analysis is dependant on the pre-processing. If a vector space representation has been chosen, the data can be analyzed using classic data-mining techniques, such as support vector machine. The vector is based on the bag-of-words model approach consisting of all words represented in the document. Clustering is an unsupervised learning problem where is necessary an automated way to organize this collection into documents relating to biomedical concepts. For the task of clustering documents the usual methods to use are unsupervised machine learning, clustering via k-means, SOM (self-organizing map) and graph based clustering. Text classification is a supervised learning problem where we know the labels of the documents (specified by domain experts) and train the corpus to effectively predict unknown future data in the right classes automatically.

[1] Available at http://medline.cos.com/
[2] Available at http://www.ncbi.nlm.nih.gov/sites/entrez

Step 4. *Visualization*: the results are graphically represented, after constructing the biological entity-document index, this it is used to compute a network connecting graphically link between every pair of genes that co-occurred.

Step 5. *Interpretation of results*: the evaluation of extracted information or validation of results. Analyzing information from biomedical text is especially challenging because of the complexity of the field. Many text-mining techniques have incorporated ontologies to take advantage of the existing knowledge that they provide.

These steps are broad research domains itself; the process of text-mining needs a well-organized integration of these phases for knowledge discovery. Since human genome sequences were first decoded, more and more researchers have become involved in this domain, especially in biology and bioinformatics. The field of bioinformatics fuses biological and biomedical sciences with information science, making possible access to vast amounts of biological information accumulated in databases. Hundreds of on-line databases characterize biological information such as sequences, structured data, and expression patterns on the one hand; and on the other, highly unstructured information in text format.

3. TEXT-MINING RESEARCH IN GENOMICS

Genomics can be said to have appeared with the initiation of genome projects for several biological species. In genomics, text-mining refers to the creation of literature networks of related bimolecular entities (Tanabe, 2005). The analysis of biomedical texts and available databases can help to interpret a phenomenon, to detect gene relations, or to establish comparisons among similar genes in different specific databases. Biological databases can generally be of two types (Stapley & Benoit, 2000): (a) biomolecular sequences and structures, such as the Swiss-Prot[3], or GenBank[4] databases; and (b) natural language text contained in databases of biomedical literature abstracts, such as Medline and PubMed. The relationship between these two forms of structured and unstructured information is key, because the literature describes essential functions of many genes. Thus, biologists can extract alignment measurements between two DNA sequences from a databank of factual resources, such as GenBank. These binary relationships can be assessed in one of two ways: as numerical values derived from alignments or co-occurrence measurements; or as symbolic values derived from semantic relations extracted from Medline.

The integration of different types of textual data in the genomic data mine will contribute towards an understanding of systems biology of different living organisms. All these processes are crucial for making sense of the immense quantity of genomic information. Text-mining research in genomics is a growing field of research involving (Tanabe, 2005):

[1] *Relationship mining* (refers to the extraction of facts regarding two or more biomedical entities).

[2] *Literature networks* (refers to the meaningful subsets of Medline based on co-occurring gene names and/or functional keywords; these the networks based on co-occurrence are motivated by the fact that functionally related genes are likely to occur in the same documents).

[3] *Knowledge Discovery in Databases* (refers to the prediction of gene function and automatic analysis of scientific papers).

Leaving that significant associations (among biological entities such as genes, proteins, and drugs) can be extracted automatically from the scientific literature, we consider below some applications that these associations have in the genome-based research, such as the genomic term identification in curation processes, the formulation of hypotheses about disease, the visualization of biological relationships, or the life-science domain mapping.

[3] Available at http://expasy.org/sprot/

[4] Available at http://www.clcbio.com/index.php?id=442

3.1 Applying Text-Mining to the Genomic Database Curation

In response to the explosion of biomedical literature, biologists develop specialized databases to organize information, such as GBD[5] *(Human Genome Database*), FlyBase[6] (*Drosophila melanogaster*), WormBase[7] (*Caenorhabditis elegans*), SGD[8] (*Saccharomyces Genome Database*), or MGI[9] (*Mouse Genome Informatics*). Researchers rely on expert-curated biological databases to organize the findings of published scientific literature. These databases collect organism genome sequences, annotate and analyze them, and provide public access. These databases may hold many species genomes, or a single model organism genome. Curators of biological databases transfer knowledge from scientific publications, a laborious and expensive manual process. Hence, curators struggling to process scientific literature need interactive tools to help transfer information from the literature into the databases (Morgan et al., 2004).

Each model organism database is maintained by a team of specialized biologists, or curators, who track the literature and transfer relevant new findings into appropriate database entries. These databases lag behind the literature because the curators have difficulty keeping up with the literature. The curation process can be separated into a series of steps (Hirschman et al., 2002): identifying new articles to be curated, reading the full-text of the selected articles and identifying the genes and/or proteins that have experimental findings associated with them, and associating functional and expression information with each gene and protein.

The curators need tools to help in the consistent transfer of information from the literature into databases. Text-mining tools can help the curators in the identification of information for genomic databases. The recent areas of text-mining for curators would be synthesized the following fields (Morgan et al., 2004): (*i*) to provide tools that can improve the currency, consistency, and completeness of biological databases; (*ii*) to explore the hypothesis that expert-curated biological databases provide resources for the creation of high quality text-mining tools that can be applied to specific curation tasks; and (*iii*) to understand the complexities of the nomenclature problem for genes.

Relating to the third issue aforementioned, numerous hurdles in genomic information are due to terminological variation and the complexity of names (Tuason et al., 2004). Irregular gene-naming arises in part because various researchers from different fields who are working on the same area of knowledge discover a large number of entities that need to be named. At present, some genes are denoted in publications under more than one name/symbol, and moreover, one symbol/name is sometimes used for several unrelated genes. There is a high correlation between the degree of term variation and the dynamic nature of genomics. As the use of gene symbols in publications can be confused approved nomenclature is intended to enable curators to access all data pertaining to a specific gene of interest, across species. Consequently, this calls for improved tools of genomic entity identification and access to full-text.

3.2 Applying Text-Mining to Form Hypotheses about Genes and Diseases

Genomic investigation takes place in highly specialized contexts with poor communication between disciples. Knowledge from one discipline may be valuable for other without researchers knowing it. As scientific publications are a condensation of the knowledge, literature-based discovery tools may help the individual scientist to explore new useful domains. Swanson (1986; 1987) was the first to make literature-based discoveries in the scientific field of biomedicine. These were later corroborated clinically and experimentally using a software system called Arrowsmith (Smalheiser & Swanson, 1998). The starting point of the proposal is the so-called '*ABC Model*': if a given concept 'A' (e.g., a disease or a gene) is associated with a second concept 'B', and 'B' is related to a third entity 'C', then 'A' might be related to 'C', even if there is no direct association between them (Swanson, 1988; Swanson et al., 2006). The initiative was adopted to establish indirect connections between *Fish Oil-Raynaud's Disease*, *Migraine-Magnesium*, and *Estrogen-Alzheimer's Disease* (Swanson, 1986; 1988; Smalheiser & Swanson, 1996).

Many other works based on statistical methods and transitive association graphs have been used for literature-based discoveries. Lindsay and Gordon (1999) developed a process that followed the same basis

[5] Available at http://www.pubgene.org/
[6] Available at http://flybase.org/
[7] Available at http://www.wormbase.org/
[8] Available at http://www.yeastgenome.org/
[9] Available at http://www.informatics.jax.org/

architecture with Arrowsmith, but they added a variety of techniques to weigh terms using information retrieval methods such as term frequency and inverse document frequency. Weeber et al. (2001) also based their work on Swanson's approach. They added both a natural language processing component to identify biomedical terms and a knowledge-based approach to help connections based on the semantic type of the connection terms; Srinivasan (2004) developed a new text-mining system called *Manjal*. She used a knowledge base for filtering terms according to their semantic types. The main difference between her system and the prior ones is that she used Medical Subject Heading (MeSH), keywords assigned to the document, to capture the content of the documents instead of applying natural language processing.

Overall, the molecular biology has moved from an era of data collection into one of hypothesis-driven means, by connecting several facts: a hypothesis can be formulated in terms that are testable by experiments (Blasoklonny & Pardee, 2002). However, because of the explosive growth in genomic literature has made it difficult for researchers to keep with advancements, while researchers formulate new hypotheses to test, it is important for them to identify connections to their work from other part of the literature. This situation offers an excellence opportunity for text-mining, i.e., to assist in the new potentially causal connections between genomic terms by automatically discovering a set of interesting hypotheses from a suitable text collection.

3.3 Applying Text-Mining to Detect Functional Relations between Genes

Leaving from the premise extensively accepted that the co-occurrence of gene terms in the same sentence or the same document often implies real biological relationships between the named entities, i.e., if two genes are co-mentioned in the biomedical literature there is an underlying biological relationship (Stapley & Benoit, 2000). The fact that a amount of biomedical knowledge is recorded in only free-text form and, as such, is not readily available for computerized analysis has inspired research on methods for automated extraction of biomedical knowledge (Andrade & Bork, 2000). Nevertheless, how best to exploit the synergies that exist between genes, sequences and texts is still an open question. The diversity of research in this area reflects the open-ended nature of the problem, although there is a common objective "to use the relationships between genes, sequences and texts as the basis for a new generation of analysis tools and methodologies that combine bioinformatics and NLP technologies" (Yandell & Majoros 2002, p. 602). Stapley and Benoit (2000) tallied the number of co-occurrences of every pair of genes in Medline abstracts and used this data to calculate what they denote as '*BioBibliometric distances*' between genes, so that the rarer the co-occurrence of two genes in the literature database, the larger the distance between them. The literature-derived gene-to-gene network may provide important information assigning a biological function to gene sequences and gene expression patterns.

Therefore, *gene-to-gene co-citation networks* can be used to test new hypotheses, and new knowledge can be generated by reviewing these accumulated results in a concept-driven manner, linking them into testable chains and networks (Jenssen et al., 2001). The nature of these relationships can be explored further using the Medical Subject Headings (MeSH®) index, or bio-ontologies. Considerable effort has likewise been centred on the construction of literature-based networks (Stephens et al., 2001; Blaschke & Valencia, 2002; Feldman et al., 2003; Krallinger et al., 2005). Novel approaches may also resort to the literature to establish functional relationships among genes, such as a methodology based on revealing coherent themes within literature through a similarity-based search in document space, after which the content relationships among abstracts are translated into functional connections among genes (Shatkay et al., 2000; 2002; Iliopoulos et al., 2001).

Ideally, all researchers would be able to associate certain related genes with others in the literature and databases. But it is difficult to know how this process is carried out. Attempts to impose standard names across the board are meeting stiff resistance, while approaches that would give genes unique ID numbers seem unlikely to take root (Pearson, 2001). Genomic is particularly dependent on shared naming conventions: if researchers cannot clearly match a name to the underlying object (gene or structure), then some failure of communication is likely to occur (Hirschman et al., 2002). Thus, this calls for improved text-mining tools of genomic entity identification and better methods for visualizing information. Building such tools is critical for managing genomic information.

3.4 Applying Text-Mining to Map the Structure of Genomic Research

There is incipient interest in learning about the structure and dynamics of the biomedical and genomic research domain by applying document co-citation networks via conceptual networks. An approach involving indexing full-text scientific articles combined with an exploratory statistical analysis may serve to complement bibliometric approaches in the mapping of science, as has been demonstrated within the biomedical field (Glenisson et al., 2003a; 2003b). The full-text analysis and bibliometric methods can be combined to improve the efficiency of individual methods in describing, understanding and visualizing the structures in a scientific field. Representations from the field of IR can be adopted for clustering of genes based on their associated literature. Similarly, there are documented attempts to develop *gene-to-literature co-citation networks* that show the interrelations among papers, genes and proteins to shed light on the structure of research regarding melanomas; thus, Boyack et al. (2004) presented an attempt to map a '*network ecology'*, that is, the interrelations among papers and genes in order to answer questions such as: *What is the structure of the research reported on a particular field? Which parts in this research field study what genomic entities? How are genomic entities and papers reporting our knowledge on them interconnected?*. The process of generating a map that shows the association linkages between papers and genes in a common context is as follows (Boyack et al. 2004): (*i*) collection of appropriate data records (papers and genes); (*ii*) calculation of pairwise similarities between records; (*iii*) layout of the records based on calculated similarities; and (*iv*) visualization and exploration of the data, enabling characterization and detail subsequently.

Text-mining can provide reliable results in representing structural aspects of bibliometric research if methods are based on full-text. Rather than discovering knowledge from data, biomedical text analysis and bibliometrics aims to give researchers a more global view of the structure and dynamics of genomic domains, to show occasions for collaboration and minimize sterile duplication of research.

4. CONCLUSION

The sequencing of the human genome has greatly increased the rate of life-science research. The biomedical literature is playing an increasingly important role in genomic discovery. The challenge is to manage the increasing volume, complexity and specialization of knowledge expressed in this literature. Text-mining tools and methods can help researchers manage this affluence of information, and discovery facts, relationships and implications in biomedical literature that can be used to assist solve genomic problems. Moreover, text-mining potentials have an increasing position to play in the broader methods of biomedical knowledge discovery, in combination with data-mining and modeling of genomic structures. As a result, there is great opportunity for applying and improving the text-mining methods in genomics outlined in this work.

REFERENCES

Andrade, M. A. and Bork, P., 2000. Automated Extraction of Information in Molecular Biology. *FEBS Letters*, Vol. 476, pp. 12-17.

Blaschke, C. and Valencia, A., 2002. The Frame-Based Module of the SUISEKI Information Extraction System. *IEEE Intelligent Systems*, Vol. 17, No. 2, pp. 14-20.

Blasoklonny, M. V. and Pardee, A. B., 2002. Conceptual Biology: Unearthing the Gems. *Nature*, Vol. 416, p. 373.

Boyack, K., Mane, K. and Börner, K., 2004. Mapping Medline Papers, Genes, and Proteins Related to Melanoma Research. *Eight International Conference on Information Visualization, Proceedings (IV'04)*. London, UK, IEEE Conference on Information Visualization, pp. 965-971.

Feldman, R., Regev, Y., Hurvitz, E. and Finkelstein-Landau, M., 2003. Mining the Biomedical Literature Using Semantic Analysis and Natural Language Processing Techniques. *Biosilico: Information Technology in Drug Discovery*, Vol. 1, No. 2, pp. 69-80.

Glenisson, P., Antal, P., Mathys, J., Moreau, Y. and De Moor, B., 2003a. Evaluation of the Vector Space Representation for Text-Based Gene Clustering. *Pacific Symposium on Biocomputing*, Vol. 8, pp. 391-402.

Glenisson, P., Mathys, J. and De Moor, B., 2003b. Meta-Clustering of Gene Expression Data and Literature-Based Information. *ACM SIG KDD Explorations, Special Issue on Microarray Data Mining*, Vol. 5, No. 2, pp.101-112.

Hearst, M., 1999. Untangling Text Data Mining. *Proceedings of ACL'99: the 37th Annual Meeting of the Association For Computational Linguistic* ACL. University of Maryland, pp. 3-10.

Hersh, W., 2005. Evaluation of Biomedical Text-Mining Systems: Lessons Learned from Information Retrieval. *Briefings in Bioinformatics*, Vol. 6, No. 4, pp. 344-356.

Hirschman, L., Park, C., Tsujii, J., Wong, L. and Wu, C. H., 2002. Accomplishments and Challenges in Literature Data Mining for Biology. *Bioinformatics*, Vol. 18, No. 12, pp. 1553-1561.

Iliopoulos I., Enright, A. J. and Ouzounis, C. A., 2001. Textquest: Document Clustering of MEDLINE Abstracts for Concept Discovery in Molecular Biology. *Pacific Symposium on Biocomputing*, Vol. 6, pp. 384-395.

Jenssen, T.-K., Laegreid, A., Komorowski, J. and Hovig, E., 2001. A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression. *Nature Genetics*, Vol. 28, No. 1, pp. 21-28.

Krallinger, M., Erhardt, R. A. A. and Valencia, A., 2005. Text-Mining Approach in Molecular Biology and Biomedicine. *Drug Discovery Today*, Vol. 10, No. 6, pp. 439-445.

Leroy, G. and Chen, H., 2005. Genescene: An Ontology-Enhanced Integration of Linguistic and Co-Occurrence Based Relations in Biomedical Texts. *Journal of the American Society for Information Science and Technology*, Vol. 56, No. 5, pp. 457-468.

Lindsay, R. K. and Gordon, M. D., 1999. Literature-Based Discovery by Lexical Statistics. *Journal of the American Society for Information Science and Technology*, Vol. 50, No. 7, pp. 574-587.

Morgan, A. A., Hirschman, L., Colosimo, M., Yeh, A. S. and Colombe, J. B., 2004. Gene Name Identification and Normalization Using a Model Organism Database. *Journal of Biomedical Informatics*, Vol. 37, pp. 396-410.

Pearson, H., 2001. Biology's Name Game. *Nature*, Vol. 411, pp. 631-632.

Shatkay, H., Edwards, S. and Boguski, M., 2002. Information Retrieval Meets Gene Analysis. *IEEE Intelligent Systems*, Vol. 17, No. 2, pp. 45-53.

Shatkay, H., Edwards, S., Wilbur, W. J. and Boguski, M., 2000. Genes, Themes and Microarrays: Using Information Retrieval for Large-Scale Gene Analysis. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*. New York, AAAI, pp. 317-328.

Smalheiser, N. R. and Swanson, D. R., 1998. Using ARROWSMITH: A Computer-Assisted Approach to Formulating and Assessing Scientific Hypotheses. *Computer Methods and Programs in Biomedicine*, Vol. 57, pp. 149-153.

Srinivasan, P. 2004. Text Mining: Generating Hypotheses From MEDLINE. *Journal of the American Society for Information Science and Technology*, Vol. 55, pp. 396-413.

Stapley, B. J. and Benoit, G., 2000. Biobibliometrics: Information Retrieval and Visualization from Co-Occurrence of Gene Names in Medline Abstracts. *Proceedings of Pacific Symposium on Biocomputing*. Hawaii, USA, pp. 529-540.

Stephens, M., Palakal, M., Mukhopadhyay, S., Raje, R. and Mostafa, J., 2001. Detecting Gene Relations from MEDLINE Abstracts. *Proceedings of Pacific Symposium on Biocomputing*. Hawaii, USA, pp. 483-496.

Swanson, D. R., 1986. Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge. *Perspectives in Biology and Medicine*, Vol. 30, No. 1, pp. 7-18.

Swanson, D. R., 1988. Migraine and Magnesium: Eleven Neglected Connections. *Perspectives in Biology and Medicine*, Vol. 31, pp. 526-557.

Swanson, D. R., 1987. Two Medical Literatures that are Logically but not Bibliographically Connected. *Journal of the American Society for Information Science*, Vol. 38, No. 4, pp. 228-233.

Swanson, D. R., Smalheiser, N. R. and Torvik, V. I., 2006. Ranking Indirect Connections in Literature-Based Discovery: the Role of Medical Subject Heading. *Journal of the American Society for Information Science and Technology*, Vol. 57, No. 11, pp. 1427-439.

Tanabe, L., 2005. The Genomic Data Mine. In: H. Chen, Fuller, S. S., Friedman, C. and Hersh, W. (Eds.), *Medical Informatics: Knowledge Management and Data Mining in Biomedicine*. New York, Springer, pp. 547-71.

Tuason, O., Chen, L., Liu, H., Blake, J. and Friedman, C., 2004. Biological Nomenclatures: A Source of Lexical Knowledge and Ambiguity. *Proceedings of the Pacific Symposium on Biocomputing*. Hawaii, USA, pp. 238-249.

Weeber, M., Klein, H., Lolkje, T. W. and De Jong-van den Berg, L. T. W., 2001. Using Concepts in Literature-Based Discovery: Simulating Swanson's Raynaud-Fish Oil and Migraine-Magnesium Discoveries. *Journal of the American Society for Information Science and Technology*, Vol. 52, No. 7, pp. 548-557.

Yandell, M. D. and Majoros, W. H., 2002. Genomics and Natural Language Processing. *Nature Reviews Genetics*, Vol. 3, pp. 601-610.