A text-mining application able to mine association rules from biomedical texts

Berardi M.(1) , Malerba D.(1) , Marinelli C.(2) , Leo P.(2) , Loglisci C.(1), Scioscia G.(2)

(1) Dipartimento di Informatica, Università degli Studi di Bari, Via E.Orabona 4, 70126 Bari – Italy (2) Java Technology Center - IBM SEMEA Sud, Via Tridente 10, 70126 Bari – Italy

Motivation

Collecting, analyzing and extracting useful information from a very large amount of biomedical texts is a difficult task for researchers in biomedicine who need to keep up with scientific advances. Nowadays several domains in medical practice, drug development, and health care require support for such actives such as bioinformatics, medical informatics, clinical genomics, and many other sectors. Moreover, for this particular task, the data to be examined (i.e. textual data) are generally unstructured as in the case of Medline abstracts and the available resources (e.g. PubMed) and as many other textual resources such as medical records, patents etc. and they do not still provide adequate mechanisms for retrieving the required information as well as to help humans in "deeply analyse" very large amount of content. In this work we present a Text-Mining framework aiming to support biomedical researchers in the task of disease-genes relationships identification from scientific abstracts retrieved by querying Medline.

Methods

Our approach exploits text-analysis technologies provided by the BioTeKS (Biological Text Knowledge Services) [5] system based on the UIMA (Unstructured Information Management Architecture) [6] both from IBM Research to identify relevant biomedical entities (e.g. genes, proteins, diseases) mentioned in a portion of a text (e.g. Medline abstracts). In particular, entity extraction performed by BioTeKS aims to both identifies the location of an entity in a text and categorize it according to the standard MeSH (Medical Subject Headings) taxonomy. A data mining technique, namely association rule mining [1], is then used to discover associations between entities as indication of the existence of a biomedical relation [3]. A discovered association rule states the co-occurrence between two sets of entities which have significant statistical evidence. Exploiting the categorization of extracted entities in the MeSH hierarchy, allows discovering association rules at multiple levels. This particular kind of association rules, namely generalized association rules [2], expresses knowledge at multiple levels of abstraction. This means that when an entity is involved in an association also its ancestor will be involved in an association regarding more general levels. Browsing and filtering methods enable biologist to navigate among subspaces of rules satisfying his/her interest as well as to select association rules satisfying domain knowledge in form of templates of rules, or statistical behaviours, or no-redundant knowledge constraints.

Results

The task of disease-genes relationships identification has been investigated by running the proposed framework on the Medline portion retrieved by means of a query on a specific disease name. Associations discovered on this set of abstracts have been used to suggest to the Medline user how to expand the initial query [4]. The expanded query allows to retrieve a new set of abstracts useful to extract gene names occurrences. This list should be considered as a

knowledge base to support the biologist in the discovery of relationships among genes and the input disease. An example of discovered association on the set of abstracts related to an initial query on therapeutic, Alzheimer(Alzheimer disease is the following: brain, amyloids disease which states the relationship among the input disease and other biomedical entities. By using knowledge expressed in this association to refine the query, the following gene names are extracted by BioTeKS in the new set of abstracts: APP, S182, STM2, SOMATOSTATI, BCL2A1, ANON, ANXA, RHO.

Contact email: berardi@di.uniba.it
Supplementary Information: Acknowledgments The University of Bari team funded in part this work with the IBM Faculty Award 2004 recently received from IBM Corporation to promote innovative, collaborative research in disciplines of mutual interest. We would like also to thank Annalisa Granatiero, Graziano Pappadà, Luigi Zanchetta and Michele Lapi for their useful contribution to this work. References 1. R. Agrawal, and R. Srikant: "Fast Algorithms for Mining Association Rules", Proceedings of the Twentieth Int.Conf. on Very Large Databases, Santiago, Cile, 1994. 2. R. Srikant and R. Agrawal: "Mining Generalized Association Rules", Proc. of the 21st Int'l Conf. on Very Large Databases, Zurich, Switzerland, Sep. 1995. 3. M. Berardi, M. Lapi, P. Leo, C. Loglisci: "Mining Generalized Association Rules on Biomedical Literature", Proc. Of IEA/AIE-2005 Eighteenth International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems, Bari, Italy, June 2005. 4. M. Berardi, M. Lapi, P. Leo, D. Malerba, C. Marinelli, and G. Scioscia. "A data mining approach to PubMed query refinement". 2nd International Workshop on Biological Data Management (BIDM 2004), in conjunction with DEXA 2004, Zaragoza, Spain, September 2, 2004. 5. R. Mack, S. Mukherjea, A. Soffer, N. Uramoto, E. Brown, A. Coden, J. Cooper, A. Inokuchi, B. Iyer, Y. Mass, H. Matsuzawa, and L. V. Subramaniam. "Text analytics for life science using the Unstructured Information Management Architecture". IBM System Journal Volume 43, Number 3, 2004 6. D. Ferrucci, and A. Lally: "UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment", Natural Language Engineering, September 2004, 10(3-4), pp. 327-348.