Construction of Gene Correlation Networks and Text Classification via Biomedical Literature Mining

George Potamias, Despoina Antonakaki, and Alexandros Kanterakis

Abstract— **Automatic extraction of information from biomedical texts appears as a necessity considering the growing of the massive amounts of the relative scientific literature. A special feature that makes this task more challenging is the over-abundance and heterogeneity of the relative genes/proteins terminology. In this paper we introduce a novel term-identification process and propose an effective data structure based on TRIE trees. It enables the storage of millions of biomedical terms and reflects their semantic relations in a compressed and memory efficient way. Gene-Gene and Gene-Disease correlations are induced based on the utilization of the entropic Mutual Information Measure. Moreover we introduce a novel texts classification process that utilizes the terms identification process and a novel similarity matching metric. The induced correlation networks reveal valuable biomedical information. Text classification results exhibit highly accuracy figures in the range of 90 to 97.5% indicating the reliability of the whole approach.**

I. Introduction

THE automatic extraction of information from biomedical texts appears as a necessity considering the growing of the massive amounts of scientific literature. The main problems are heterogeneity of used *vocabularies* and *lexical coverage*. The problem arises from the fact that there is not a standard adopted *terminology*. The emerging need is organization and centralization of the different biomedical terminological references, a task that calls experts from different but eventually assembled sections of science. In this context the use of pronouns and definite articles, long, complex or negative sentences or, those in which information is implicit can be also inconvenient for a searching algorithm. *Term ambiguity* can arise from the identification with common English words or bad encoding of genes and proteins (in the rest of the paper 'gene' and 'protein' are used interchangingly).

Literature data mining is concerned mainly with the discovery of valid, novel, interesting patterns, associations and deviations in scientific literature [1], [2], [3]. It comprises two technologies: *information extraction* and *text mining*. Information extraction concerns the task of identification and extraction of the relevant information from the accumulated texts, according to user's requests. Text mining is defined as the process of discovering and extracting knowledge from unstructured data, contrasting it with data mining which discovers knowledge from structured data [4].

Referring to *Term Identification* we should decompose it in three major steps: term *recognition*, term *classification*, and term *mapping*. In this paper we are mainly concerned with term recognition which refers to the marking of the words belonging to the domain, in the literature (in the rest of the paper 'term identification' and 'tern recognition' are used interchangingly). The occurrence of a single term has such significance as well as the co-occurrence with other terms. Potential considerations that must be consulted are the differentiation between terms and 'non-terms' (i.e., terms and lexigographic entries with no direct semantic relation to the target biomedical domain), and the variation of a specific one.

Machine learning systems as well as statistical techniques are thoroughly used to cope with these problems. General machine learning approaches usually used include: decision tree learners, neural networks and support vector machines as in [5] and in [6], [7]; (Naïve) Bayesian approaches as in [8]. In [9] a Hidden Markov Model (HMM) approach, coupled with specific orthographic features, is utilised for the discovery of terms from a set of ten classes (a recall/precision F-score of 75.9% is reported). Similar results, for the recognition of Drosophila gene names using also HMMs, are reported in [10].

Another trend used towards term identification is based on *hybrid* approaches and systems. They combine rule-based approaches, with statistical techniques as well as linguistic and contextual processing in order to rank candidate terms. A protein/gene name tagger, ABGene, is presented in [11] – it was trained on Medline abstracts by adapting a POS (part of speech) tagger achieving a precision in the range of 60% to 90%. Another remarkable hybrid method called "C/NC value" is presented in [12] – experimental results on 2,082 MEDLINE abstracts showed a precision of 91-98% for the top ranked terms. In [13], term recognition is based on a scheme that supplements sequence similarity. In [14] a method is proposed for clustering abstracts based on a statistical treatment of terms, together with stemming, a 'go-list', and unsupervised machine learning.

Manuscript received on June 30, 2006.

All authors are within the Institute of Computer Science, FORTH, Heraklion, Crete, Greece (Corresponding author: George Potamias, phone: 30-2810-391693; fax: 30-2810-391601; e-mail: potamias@ics.forth.gr).

A primary challenge for our work is the task of efficient retrieval of data in linear time based in the implementation of a *Trie* memory-based structure [16], [17]. The data include a set of *stop-word* dictionary, biomedical *terms*, and *free text* descriptions of terms.

A primary challenge for our work is the task of efficient retrieval of data in linear time based in the implementation of a *Trie* memory-based structure [16], [17]. The data include a set of *stop-word* dictionary, biomedical *terms*, and *free text* descriptions of terms. A special prerequisite is to capture, represent and record the *semantic interrelation* between terms as well.

Unsupervised machine learning techniques are appropriately utilized in order to induce reliable Gene-Gene and Gene-Disease *correlations* contained in the biomedical literature. *Mutual Information Measure* (MIM), a well established entropic metric, is used in order to estimate the strength of induced correlations. For each correlation, the pair of the gene terms, the abstracts in which they are identified, and disease nomenclature are appropriately utilized. The most expressive relations are discovered and a *terms correlation network* is constructed and visualized.

Moreover, we introduce and present an efficient and reliable approach for the classification of text-references (i.e., Pubmed abstracts) based on the identified terms and a novel similarity matching formula.

We have implemented all the relevant operations in a system called *MineBioText* [15]. A general view of the system's architecture is shown in Fig.1, below.

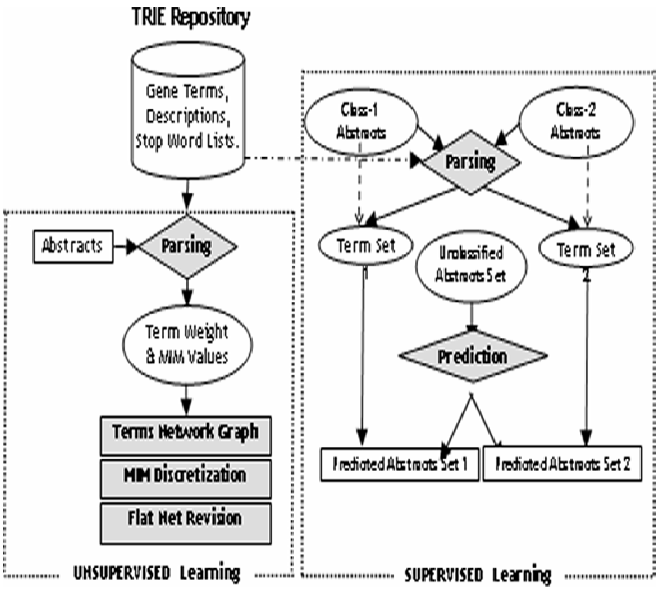

Fig. 1. The general architecture of the MineBioText system. In unsupervised learning mode the input terms and abstracts are passed through the post processing phase (Parsing). The output graph visualization includes the gene terms or/and the diseases, according to the significant. In supervised learning mode the input terms and abstracts are also passed through the post processing phase. The output, apart from the predicted classes of the abstracts, includes the accuracy of the prediction and the AUC/ROC estimation.

II. METHODS

A. The Input

Initially a corpus of relevant biomedical text references is collected, i.e., *PubMed* abstracts. With the aid of *Ensembl*'s *BioMart* (www.ensembl.org/Multi/martview) - a data mining tool that can be used with any type of data and provides a build-in support for query optimization, we accumulated the respective gene/protein terminology. BioMart provides a set of filters in order to include or exclude characteristics of the retrieved gene/protein names. We also retrieved human gene terms from the gene ontology - GO database (www.geneontology.org) which provides a *free-text description* for each gene/protein. Ensembl identifier was used as our *primary* gene/protein reference *identifier*.

B. The Trie Data-Structure and its Utilization

A common issue in data mining is the demanding and overwhelming amount of data. Considering the potential overhead of a database usage, we tried to take advantage the efficient indexing methods used in database systems. An eventual relief was the adjustment of the *Trie* memory in the domain [16], [17]. Concerning speed, memory need, and sensitivity of parameters Tries were proven to outperform hash-trees [18].

We have employed and implemented a double-chained *Trie* where, the edges of a node are stored in a double connected list. An example of the way this structure is utilised for terms' storage is illustrated in Fig. 2.

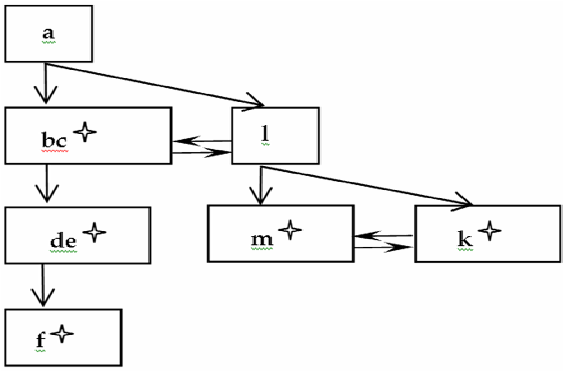

Fig. 2. An example of how terms are stored**.** The tree contains terms "abc", "abcde", "abcdef", "alm", and "alk". Each node contains a unique symbol or group. Each leaf of the tree which contains '✧', is a complete term. A term is composed of all the above ancestors until the root of the tree. All the other nodes simply represent a common symbol of other terms.

As a more demonstrative example, for the case of inserting a term into the structure, assume that the trie holds ENSG00000135487 (the Ensembl Gene ID) and HLMKL2 (a gene in HUGO terminological notation), and that ENSG00000135486 and HNRPA1 are to be inserted. Fig. 3, below, shows the state after the insertion where the bold lines indicate the links that connects the new inserted gene terms.

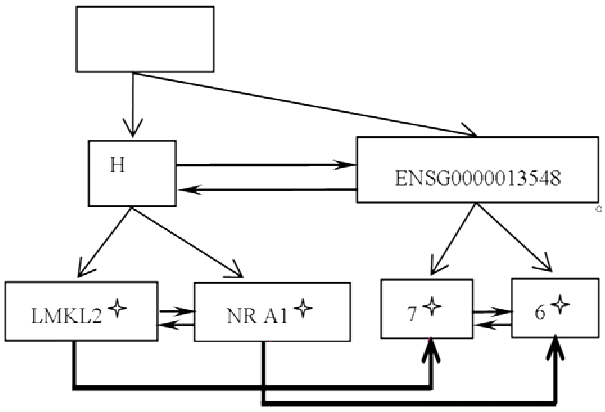

Fig. 3. Asserting and retrieving gene/protein references into the trie structure.

The time complexity for the insertion and retrieval of terms with our Trie-structure implementation can be assessed assuming a word of *n* letters. The search process will seek for the first letter in all the nodes of the built tree, in order to figure out its ancestor. The time complexity of this action depends only on the amount of letters contained in the tree, suppose *c*. Each term of n letters will take *c*n* steps, so the total search time complexity is *O(c*n)*. It can be proven that in an implementation without a double connected list this time can be reduce to *n*logc*.

- *Text Parsing.* The first problem in parsing free text references of biomedical content is the removal of *string patterns* that contain *common words* (i.e., words with no semantic relation to the target biomedical domain). An efficient way to cope with this problem is to eliminate pre-specified patterns by using list of *common words*, and employing a look-up approach. A dictionary of English common used words is utilized for this purpose (http://wordlist.sourceforge.net/). Finally parsing of the gene/protein terms is necessary in order to increase sensitivity and reduce parsing time. Note that gene/protein names and symbols are converted into lowercase; with punctuation marks and others symbols removed. As long as the parsing process searches for single-terms a stemming operation is needless, i.e., potential common words within the text will never be reached following the *Trie*-based search process presented above. Note that with the MineBioText system the user may customize (e.g., adding/removing words and phrases) the exclusion common-words dictionary to meet her/his needs.
- *The utilized standard gene/protein terminologies.* The process of the localization and recognition of terms utilizes various sources of gene/protein terminologies and nomenclatures. Primarily the inquiry is based on the combination of the 'Ensembl Gene ID' – as the *primary gene/protein reference key*, as well as other identifiers utilized as standard gene/protein *synonyms*. These gene/protein synonyms are provided from various related gene/protein nomenclature resources and gene/protein namings: 'GO Id' and '*GO descriptions'* (www.geneontology.org), 'HUGO id' (http://www.gene.ucl.ac.uk/nomenclature/), 'OMIM id' (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM), 'Uniprot Swissprot id' (http://www.ebi.uniprot.org/index.shtml). All these nomenclatures are appropriately incorporated and utilized by the MineBioText system with the respective gene/protein term references being searched in all input text references.
- *Gene/Protein Identification and GO-descriptions.* In order to utilize free-text (i.e., GO) descriptions we need to find and register respective *gene/protein lexicographic-identifiers* – '*gli'*, which are *descriptive* of (*relates* to*)* specific gene/protein terms. It is a process that extends the genes/proteins terms with *extra synonyms*. Here we are faced with the problem of words (or, roots of words) contained in many gene/protein descriptions. To cope with this problem we follow an intelligent parsing operation of the GO free-text descriptions in order to assess and measure the *degree* of *gli relevance – gli_r* of the description-words with respect to the corresponding genes/proteins. We cope with two cases: (i) if the *gli* is found in just one single description then its *gli_r* is set to *1*; and (ii) if the *gli* is located in more than one description, its *gli_r* is computed by the sum of all the previous calculated weight values for it ($SUM_{other\text{-}gli\_r}$), plus 1 divided by the total number of descriptions where the *gli* is found ($Description_{gli}$). Note that we may result into *lgi_r* values that are greater than *1*. In a more formal setting: $gli\_r = SUM_{other\_gli\_r} + (1 / Description_{gli})$. The parsing process and the above formula present a form of *term normalization*.

C. Weighted Vector-based Representation

We employ a *vector-based* approach to register the occurrences of every term in the input texts. We deviate from the classic binary vector-based representation and move towards a more 'vague' assessment and registration of identified genes/proteins. During parsing the located words should be tested for their *relevance* with respective genes/proteins. For this purpose, a special process is devised and implemented. It copes with two cases: (i) the located word matches a word in a GO description - its weigh is set equal to its respective (computed and recorded) *gli_r* - note that in this case *gli_r* values may be greater that 1; and (ii) the located word matches a gene/protein term - its weight value is assigned to the largest weight value from all other identified words in the text (also taking into account the previous case).

We introduce the following notation for each of the input set:

- *Abstracts*. We define an abstract as a_i that belong to A and a_i a subset of Λ where, Λ is a potential set of words, and $\forall a_i \in A$ where $a_i \subset \Lambda$, $a_i = \{\lambda_{i1}, \ldots, \lambda_{iki}\}$ and $k_i = |a_i|$ the size of a_i.
- *Set-of-Abstracts*. Assume $A = \{a_i \ldots a_n\}$ as the finite set of the abstracts. The total number of abstracts is denoted as $|A|$.
- *Set-of-Terminology-Terms*. We denote the set of all terms from the utilized gene/protein nomenclatures as T_{nom}; with different instantiations for each of the respective gene/protein nomenclatures, e.g., T_{HUGO} for HUGO, $T_{UniProt}$ for UniProt, $T_{SwissProt}$ for SwissProt etc. A single gene/protein terms is denoted with t_x.
- *Set-of-All-Terms*. We define the set of all terms – apart from the Ensembl identifiers- as T_X; with $T_X = T_{HUGO} \cup T_{EMBL} \cup \ldots$
- *Set-of-Ensembl-Terms*. We denote the set of the Ensembl identifiers as $S = \{s_1 \ldots s_m\}$; the size of S is denoted as $|S|$.
- *Description* (free-text). A *description* $\boldsymbol{t_D}$ is a set of words, and is defined as: $\forall t_D \in T_D, t_D \subset \Lambda$ $t_D = \{\lambda_1 \ldots \lambda_N\}, N = |t_d|$
- Set-of-*Descriptions (lgi)*. A set of descriptions T_D is the set of all t_D defined as: $t_D \in T_D$ *where* t_D *is a set of words* Λ^{κ}
- Set-of-*Common-Exclusion-Words – the Words List.* Is denoted with L, L = *{the set of all English words in the input common-words file}*.

Initially all the gene/protein terms T_D and S are stored. Assume that during parsing, and for each gene/protein contained and located in set T_X, as well as the terms contained in the descriptions T_D, the Ensembl primary gene identifier is located, selected, and its *significance is* estimated. The *significance* of a gene/protein term is defined as a function: $\forall t_x \in T_x$, $\exists s_{tx} \in S$, *such as:* $\exists T \rightarrow S$. For each GO-description there is a set of *significant identifiers* (i.e., *gli*) S_{tD} that belongs to S: $\forall t_D \in T_D$ *corresponds an* $S_{TD} \in S$.

Computing terms weights. Equipped with the definitions made above, assume a word being located in a text reference. If the word belongs to the set of standard gene/protein terminological references (Tx) or, to the Ensembl's identifiers set (Sx), the weight value assigned to its corresponding significant identifier is set to 1. Otherwise, we check if it belongs to the set of descriptions; if it belongs to one description we set its corresponding Ensembl identifier equal to 1; if the word belongs to more that one description (i.e., it corresponds to different genes), assume n, its significant identifier weight is set to $1/n$. Finally all weight values are pruned to one. At the end, and for each text reference, its respective *weighted-vector* representation is formed. The number of places of the vector is always fixed and equal to the total number of Ensembl identifiers, with their computed weights as values.

D. Construction of Gene/Protein Associations Network

In order to identify the terms that share implicit associations, a knowledge association's network is build using statistical techniques. *Mutual Information Measure* (MIM), an entropic measure, is utilized in order to quantify *correlations* between variables, based on co-occurrence statistics [19], [20]. MIM computes the *correlation* or, *association strength* between gene/protein terms with reference to a given collection of abstracts (Fig.4).

$$\mathrm{MIM}(i, j) = \sum_{k=0,1} P_{k;i} \times P_{k;j} \times \log \frac{P_{k;i,k;j}}{P_{k;i} \times P_{k;j}}$$

Fig. 4. *Mutual Information Measure*: $P_{k;i}$ denotes the percentage of input abstracts in which term i occurs (k=1) or do not occurs (k=0). $P_{k;i,kj}$ denotes the percentage of abstracts in which terms i and j co-occur (k=0) or, none of them occur (k=0).

Previous work has shown that it is possible to identify implicit relationships by ranking inferred relationships and preferentially examining those at the top list [21]. The computed MIMs are stored in a file to be used for the construction of gene/protein correlation (or, association) network (presented into the sequel).

Forming the Gene/Protein Correlations Network. The next step includes the construction of the *genes/proteins correlation network*. It is based on the appropriate elaboration of the computed gene/protein MIM values. The whole process follows two steps: (i) initially the list of terms; the list of abstracts; and a user specified *percentage MIM-threshold* for the gene/protein terms with top ranked MIM values, are provided. The last input specification is provided in order to *filter-out* the gene/protein MIM values that are below the specified MIM threshold. This is done in order to keep the *most-informative gene/protein correlations*; and (ii) after filtering-out, the remaining gene/protein correlations are also examined for their *strength*. This operation is performed with a careful *discretization* of the corresponding MIM values into three correlation strength levels, with the following natural interpretation: *strong*, *medium* and *weak*. Discretization of MIM values is based on a method reported in [22], also utilised in [23] in the context of gene selection.

E. Classification of Texts

We introduce a novel approach for *text categorization* and text *class/category-prediction* based on term frequency and *supervised learning* approaches.

Training-phase. For simplicity of the presentation, assume a two-class (categories) problem - *POS* and *NEG*.

The process may be generalized to cover multi-class cases. Two corresponding sets of abstracts are assumed to be available - by querying Pubmed for specific categories of interests (e.g., 'breast cancer' vs. 'ovarian cancer'). Training is performed on each of the class-specific set of abstracts. The corresponding abstracts are parsed and for each abstract, its corresponding vector-based representation is formed. Then, weight computation is performed according to the methodology presented above. The *strength* values for the significant identifiers is computed as the sum of the weight values of all the terms identified (formula in Fig. 5, above). All the results are stored into the *train-results-file.*

$$s_{A;train} = \sum_{i=1\ldots L} Vt_{A;train_i}$$

Fig. 5. *Calculation of Strength Values. A;train* is the set of training abstracts; L is the number of gene terms located in *A;train*, $t_{A;train}$ is the set of gene terms located in *A;train*, $Vt_{A;train}$ the Ensembl unique identifier of each gene term, and $S_{A;train}$ the strength of the identifier (to compute).

Testing-phase. As for the training case, we assume the availability of two class-specific sets of abstracts. For each set the identified gene/protein terms and their weights are recorded and stored. For each term identified in the set of test abstracts its occurrence in the saved train-results-files is checked, as well as its corresponding class-specific *rank*, i.e., its position in the ordered (by their training strengths) lists of the corresponding file. So, we have different ranks for the POS, $rank_{POS}(t)$, and for the NEG, $rank_{NEG}(t)$ classes, respectively. The formula in Fig.6, below, computes the *strength,* $strength_{TEST}(t)$, of a term t identified in a test-abstract. It is computed with reference to its weight, *weight(t)*, and its corresponding and class-specific strengths, $strength_{POS}(t)$ and $strength_{NEG}(t)$.

$$strength_{TEST}(t) =$$

$$\frac{rank_{TRAIN\_POS}(t)}{count_{TRAIN\_POS}} - \frac{rank_{TRAIN\_NEG}(t)}{count_{TRAIN\_NEG}} \times weight_{TRAIN}$$

$$\times \left| \frac{strength_{POS}(t)}{max(strength_{POS})} - \frac{strength_{NEG}(t)}{max(strength_{NEG})} \right|$$

Fig. 6. Computing the strength of a term in a test text-abstract.

For all the terms identified in a test abstract the sum of their corresponding strengths is computed. If it is greater than *0*, it is assigned (classified) to the *POS* class, otherwise to the *NEG* class.

III. EXPERIMENTS AND RESULTS

In order to evaluate the reliability of the presented biomedical literature mining approach, we focused on five domain sets including retrieved sets of abstracts and gene/protein terms from PubMed and Ensembl that concern the '*Colon'*, '*Breast'*, '*Leukaemia*', '*Ovarian'* and '*Prostate' cancer* domains. The sets of abstracts were compared in order to exclude the common abstracts (i.e., the abstracts that refer to more than a single domain of interest).

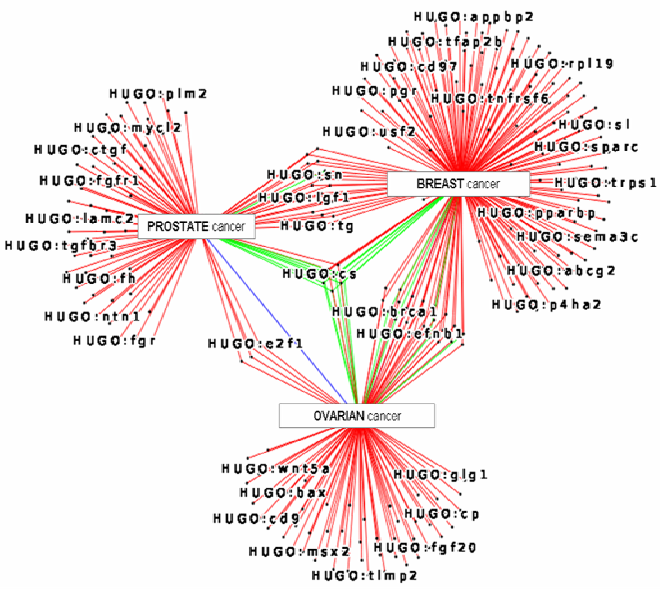

Fig. 7. The visualized gene/proteins-disease correlation network between Breast, Ovarian and Prostate Cancer – 'red', 'green' and 'blue' links indicate 'strong', 'medium', and 'low' correlation strengths, respectively. Visualization of the network is achieved with the use of the TULIP graph visualization tool (http://www.tulip-software.org/).

Correlation Networks. Gene and Gene-Disease correlation networks were generated for all the domains - Fig. 7, below, illustrates such a network. It concerns the breast, ovarian and prostate cancer cases. Inspecting the network one may identify genes/proteins that correlate with a specific disease or, identify genes/proteins common to two diseases (e.g., gene 'brca1' is common between the breast and ovarian cancer).

Texts/Abstracts-Classification. A number of 9258, 4594, and 13218 abstracts were retrieved for breast, colon, and leukemia diseases, respectively (by querying Pubmed and excluding the common, between diseases, abstracts). Classification results, based on a 50% split of the abstracts to training and test sets, are shown in table I, below.

TABLE I
TEXTS/ABSTRACTS CLASSIFICATION RESULTS

Task	Accuracy %	ROC/ AUC
'*Breast*' *vs* *'Colon'*	93.0	0.993
'Colon' vs 'Leukemia'	97.5	0.997
'Breast' vs. 'Leukemia'	90.0	0.966

The results – both in terms of accuracy and ROC/AUC figures, are indicative for the reliability of the whole gene terms' identification and weighting process, as well as for the introduced texts classification metric.

IV. CONCLUSIONS AND FUTURE R&D PLANS

We presented an integrated biomedical literature mining methodology for the identification of gene/protein terms, assessment of their relative (to each text reference) weights and strengths. Although the heterogeneity and complexity of terminology and nomenclature in the biomedical domain, we presented an effective approach for the distillation of valuable information, as being exhibited from the construction of gene and gene-disease correlation networks.

The overall methodology is based on: (a) the introduction of an efficient and effective memory structure (the Trie data-structure) - able to cope with the huge amounts and the semantic heterogeneity of the involved biomedical nomenclatures; (b) the introduction of special term-weighting metrics; (c) the utilization of the mutual information measure to assess the correlation strength between two gene/proteins, and the subsequent construction of the correlation networks; and (d) a specially devised text classification process and related metrics.

The overall methodology is implemented in an integrated (and easily adaptable to different domains) system called MineBioText. We have tested and examined the performance of the system on various biomedical domains achieving very good accuracy and sensitivity/specificity figures.

Intensive experimentation with different biomedical domains – to test the reliability and efficiency of the system, as well as porting of the whole system into a Web-services (and application) environment compose our future R&D targets.

ACKNOWLEDGMENT

The work was partly supported by the INFOBIOMED NoE FP6-IST-2002-507585 and ACGT FP6-IST-2005-026996, EU funded projects. Opinions and results expressed herein do not correspond to official projects' Consortium position and are the sole responsibility of the authors.

REFERENCES

[1] R. Feldman, "Mining unstructured data", *KDD Tutorial Notes*, pp.182–236, 1999.

[2] D. Mladenic, "PhD thesis", http://www-ai.ijs.si/DunjaMladenic/PhD.html, 1998.

[3] F. Ciravegna, "Challenges in information extraction from text for knowledge management", in IEEE Intelligent Systems and Their Applications, (Trend and Controversies), 2001.

[4] M. A. Hearst, "Untangling text data mining", in Proc. 37th Annual Meeting of the Association for Computational Linguistics, pp. 3–10, 1999.

[5] B. Stapley, L. Kelley, and M. Sternberg, "Prediction in the sub-cellular location of proteins from text using support vector machines", in Procs of the Pacific Symposium on Bio-Computing – PSB, pp. 374-385, 2002.

[6] R. Bunescu, R. Ge, and R. J. Mooney, "Extracting gene and protein names from biomedical abstracts", unpublished technical note, http://www.cs.utexas.edu/users/ml/publication/ie.html, 2002.

[7] R. Bunescu, R. Ge, R. J. Kate, R. J. Mooney, and Y. M. Wong, "Learning to extract proteins and their interactions from MEDLINE abstracts", in Proceedings of the ACM symposium on Applied computing, pp. 121-127, 2004

[8] K. S. Sathiya, *et al.*, "A machine learning approach for the curation of biomedical literature", *ACM SIGKDD Explorations Newsletter*, vol. 4, no. 2, pp. 93-94, 2002.

[9] N. Collier, C. Nobata, and J. Tsujii, "Extracting the names of genes and gene products with a hidden markov model", in Proceedings of COLING, Saarbruecken, 2000, p. 201–207.

[10] A. Morgan, A. Yeh, L. Hirschman, and M. Colosimo, "Gene name extraction using Fly Base resources", in 2003 Proceedings of NLP in Biomedicine - ACL, Sapporo, Japan, 2003, p. 1–8.

[11] L. Tanabe, and W. J. Wilbur, "Tagging gene and protein names in biomedical text", *Bioinformatics*, 2002. vol. 18, no. 8, pp. 1124–1132, 2002.

[12] K. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms: the C value/NC-value method", *International Journal on Digital Libraries*, vol. 3, no. 2, pp. 115-130, 2000.

[13] J. T. Chang, S. Raychaudhuri, and R. B. Altman, "Improving biological literature improves homology search", in Pacific Symposium on Bio-computing, Mauna Lani, 2001, HI, 374-383, 2001.

[14] I. Iliopoulos, A. Enright, C. Ouzounis, "Textquest: document clustering of Medline abstracts for concept discovery in molecular biology", in Pac. Symp. Biocomput., pp. 384–95, 2001.

[15] D. Antonakaki, A. Kanterakis and G. Potamias, "Biomedical literature mining for text classification and construction of gene networks", in Proceedings of the 4th Hellenic Conference on Artificial Intelligence, *Lecture Notes in Computer Science - LNAI* 3955, pp. 469-473, 2006.

[16] V. Aho, J. E. Hopcroft and J. D. Ullman, *Data Structures and Algorithms*, Addison-Wesley, Reading,Mass., 1983, pp. 163–169.

[17] E. Fredkin, "Trie memory", Informal Memorandum. Bolt Beranek and Newman Inc.. Cambridge, Mass., 23 January 1959.

[18] F. Bodon and L. Ronyai, "Trie: an alternative data structure for data mining algorithms", *Computers and Mathematics with Applications*, vol. 38, no. 7, pp. , 739–751, 2003.

[19] B. J. Stapley, and G. Benoit, "Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts",. in Pacific Symposium on Bio-computing - PSB, pp. 529–540, 2000.

[20] T. Dunning, "Accurate methods for the statistics of surprise and coincidence", *Computational Linguistics*, vol. 19, pp. 61–74, 1993.

[21] J. D. Wren, R. Bekeredjian, J. A. Stewart, R. V. Shohet, and H. R. Garner, "Knowledge discovery by automated identification and ranking of implicit relationships", *Bioinformatics*, vol. 20, pp. 389–398, 2004.

[22] L. M. Lopez, I. F. Ruiz, R. M. Bueno and G. T. Ruiz, "Dynamic discretisation of continuous values from time series", in R.L. Mantaras and E. Plaza (Eds) Proc. 11th European Conference on Machine Learning (ECML 2000), *LNAI* 1810, pp. 290–291, 2000.

[23] G. Potamias, L. Koumakis, and V. Moustakis, Gene selection via discretized gene-expression profiles and greedy feature-elimination. *LNAI* 3025, pp. 256–266, 2004.