Exploring the correlation of biomedical article keywords to MeSH terms

Theodosios Theodosiou, Lefteris Angelis and Athina Vakali

Abstract — **The exponential growth in the availability of biomedical information has posed the need to solve retrieval issues raised in huge sequence/biomedical article repositories. Biomedical article databases, like PubMed, are huge repositories of useful biological information given in natural language form and thus not easily processed by computers. Medical Subject Headings (MeSH) terms have been proposed to facilitate the process of electronically retrieving biomedical articles, which are semantically related. However, most of the classification algorithms, used for information retrieval, require numeric representations of either the keywords or the MeSH terms of the articles. These representations are essentially vectors of variables forming large multivariate numerical datasets. In order to combine the information from keyword datasets and MeSH datasets, this paper proposes a multivariate statistical approach which can quantify their relationships and reveal the underlying correlation. The basis of this approach is a mathematical technique, called non-linear canonical correlation analysis (NLCCA). NLCCA can assemble information from several datasets by building a model describing the whole of the data. The method was applied to a large number of articles from PubMed. Certain statistics obtained from the analysis showed that the degree of correlation between MeSH terms and keywords is high. The method results in the reduction of data dimensionality, containing in one dataset with new variables significant information of the original data. These results are very important for the efficient description and visualization of the data in order to explore their structure.**

I. Introduction

The use of high-throughput experimental methods, such as microarrays, and the results from large-scale computational experiments produce huge amounts of biological information [1], which is usually stored in data repositories, like sequence databases. One of the biggest repositories of biomedical information, used extensively by researchers, is the PubMed database [2]. PubMed is a database of published biomedical articles containing bibliographic citations and abstract articles from more than 4800 biomedical journals [3].

As biological information increases its retrieval from the databases becomes a quite difficult task since the information is stored in natural language form, which can not be easily processed by computers [2]. To facilitate this retrieval, there are various efforts based on metadata, like the Medical Subject Headings (MeSH)[1] terms [4].

The MeSH terms are metadata in the sense that they provide a consistent way of representing the concepts described in biomedical articles. They are part of a standardized biomedical vocabulary that has a specific structure. The terms are arranged both alphabetically and hierarchically. The most general levels of the hierarchical structure contain the most general MeSH terms, whereas more specific terms are to be found in narrower levels. The main use of MeSH terms is to enhance the search and improve the accuracy of the query system of PubMed [4], [5]. Each PubMed article can have more than one MeSH term assigned to it, usually 10-12, while some of them describe the major topics of the article [4]. The assignment of each MeSH term to an article is done by experts categorizing biomedical articles.

Since the MeSH terms describe the content and main concepts of an article, it is reasonable to assume that they are highly correlated to the article keywords. The problem is that although we understand this semantic relationship, it is not easy to have a numerical indication of "how much" correlated are two sets of data describing the same biomedical articles, the first by their keywords and the second by the MeSH terms assigned to it. So, the main idea behind our work is to provide a mathematical framework for the quantification of such correlation.

The MeSH terms have been studied in earlier efforts mainly with respect to the retrieval of biomedical articles in order to assess their usefulness as metadata [5], [6], [7]. They have also been used with success for text classification and categorization [8]. A common characteristic of all the research efforts is the data representation, based on the Vector Space Model (VSM) [9]. VSM transforms a document into a vector of weighted keywords or MeSH terms, suitable for statistical analysis. However, these vector representations are used by the various data mining and machine learning techniques without any consideration for quantifying or interpreting the underlying correlation between the vectors.

The motivation for the present work is to investigate statistically in what degree the semantic relationship between the keywords of the biomedical articles and the MeSH terms is preserved between their vector representations. For the purposes of this analysis a multivariate statistical method, called non-linear canonical correlation analysis (NLCCA), with optimal scaling [10] was used. The method's general

Manuscript received June 30, 2006.

All authors are with the Department of Informatics, Aristotle University of Thessaloniki, 54124, Greece. (corresponding author T. Theodosiou phone: +302310-998236; e-mail: theodos@ csd.auth.gr).

[1] http://www.nlm.nih.gov/mesh/meshhome.html

goal is to explain as much as possible of the variance in the relationships among the variables of the original datasets in a low dimensional Euclidean space. The standard canonical correlation analysis (CCA) is a linear multivariate statistical method which is considered to be an extension of multiple regression and is used for two datasets of usually numeric variables. On the other hand NLCCA is not restricted to linear relationships and is especially useful for analyzing all kinds of data, numerical and categorical. Moreover, it can be applied to more than two datasets.

In general, the advantages of applying the NLCCA method to datasets obtained from biomedical articles can be advantageous in many ways. More specifically, we can:

- Assess the usefulness of the metadata as proper descriptors of the information contained in the biomedical articles.
- Explore the importance of each metadata term for describing the data.
- Combine information from the keywords of the articles and their relevant MeSH terms in order to produce a model that will describe the articles with one dataset with much less variables containing information from both the keywords and the MeSH terms.
- Reduce the number of variables (data reduction) for describing the information contained in the articles.
- Visualize the data in a low dimensional Euclidean space.
- Discover certain patterns, groupings, trends and special cases in the graph plots of the data.

In the following section we describe the general principles of the NLCCA method. In section 3 we present the results of applying the method to our data. Finally, in section 4 we provide a conclusion and directions for future work.

II. Methodology

A. Data Representation

The source of published biomedical literature is the PubMed database. The articles[2] in this study are represented by the use of VSM. The steps of transforming each article into a vector are described briefly below [11]:

1. **Tokenization**: extraction of all words appearing in an entire set of articles;
2. **Stopword removal**: elimination of non informative words (stopwords) such as ”a”, ”and”, ”the”, etc.;
3. **Stemming**: use only the root of each word;
4. **Frequency counting**: counting of the number of occurrences of each word in each document;
5. **Filtering**: elimination of non-content-bearing high-frequency and low-frequency words;
6. **Vector creation**: construction of a weight vector. In our study we used the binary weighting scheme, which is the simplest. According to it, the weights are binary numbers (0 denoting the absence of a word and 1 denoting the presence).

We have to emphasize, that the statistical methodology we used can be applied as well to other weighting schemes appearing in the literature [11] with real-valued vectors.

Regarding the MeSH terms, only steps 4 and 6 are used, since MeSH terms are already provided for each article by the PubMed system. The aforementioned procedure results in two different datasets corresponding to the same body of articles: The keyword dataset (denoted by $\mathbf{A}$) where each article is represented by a binary vector of size p (total number of keywords) and the MeSH term dataset (denoted by $\mathbf{B}$), of exactly the same articles, where each article is represented by a binary vector of size q (total number of MeSH terms).

B. The NLCCA method

The Non-Linear Canonical Correlation Analysis (NLCCA) is a multivariate statistical technique aiming to explore and model the strength of the correlation between two or more datasets. NLCCA is based on a mathematical method called optimal scaling, which reduces the original number of variables and scales the data in order to project them in a low-dimensional Euclidean space [10].

NLCCA is applied to multidimensional numerical matrices, like the binary matrices $\mathbf{A}$ and $\mathbf{B}$ that we described earlier and quantifies the correlations between them. The main and the most interesting feature of the method is that it produces a new matrix with much fewer real-valued variables, the so-called object scores (in our case we can call them article scores). These new variables form the low-dimensional Euclidean space where the articles are now projected as points, preserve a large part of the information contained in both the keywords and the MeSH terms.

This type of projection in the form of points is particularly useful for visualization of the data and exploration of patterns and groupings. Intuitively, we expect that relevant articles with similar keywords and MeSH terms will be represented by points closely positioned in the new Euclidean space, while irrelevant articles will be far away. Another appealing aspect of the method is that the resulting variables are ranked according to their importance for explaining the overall correlation. So, the first score contains the most significant part of the information from both data sets, the second score contains the second significant part and so on. This property is very useful in the sense that we can draw important conclusions regarding the data structure by plotting the articles in only two or three dimensions defined by the first scores.

The mathematical background of NLCCA with optimal scaling involves the numerical solution of an optimization problem. The mathematical problem can be described as the minimization of the amount of information that is lost when we construct a low dimensional map of the articles, the keywords and the MeSH terms in Euclidean space. The

[2] In our context an article signifies the title and the abstract only and not the whole document.

construction of such a map involves the fitting of a model which has to be the best among all possible ones.

The optimality of the model can be achieved by the minimization of a loss function defined as:

$$\sigma(\mathbf{X};\mathbf{Y_1},\ldots,\mathbf{Y}_J) = K^{-1}\sum_{k=1}^{K} tr\left[\left(\mathbf{X} - \sum_{j\in J(k)} \mathbf{G_j Y_j}\right)'\left(\mathbf{X} - \sum_{j\in J(k)} \mathbf{G_j Y_j}\right)\right] \quad (1)$$

where in general, by K we denote the number of datasets and by $J(k)$ the number of variables in k-th dataset. The $\mathbf{G}_j$ are (0,1)-matrices each one corresponding to a variable, with elements indicating the existence or the absence of a category (value of the variable) in the j-th variable $\mathbf{v}_j$. In our case where all the variables (keywords and MeSH terms) are binary (values 0 or 1), $\mathbf{G}_j = [1 - \mathbf{v}_j, \mathbf{v}_j]$. By $tr(\mathbf{H})$ we denote the sum of the diagonal elements of a matrix $\mathbf{H}$.

So, the objective is to estimate the matrices $\mathbf{X}$ and $\mathbf{Y}_j$, $j = 1,\ldots,J(k)$, $k = 1,..,K$ which simultaneously attain the minimum value of the loss function. Note that in our setup $K = 2$, $J(1) = p$ and $J(2) = q$.

The above problem can be solved computationally efficiently by the Alternating Least Squares algorithm [10]. The algorithm is implemented in the SPSS statistical software which we use for the purposes of our study [12].

The output of the NLCCA method contains a series of measures (Table I) that depict the degree of correlation between the datasets. Also, the method calculates the efficiency of the model as a whole and the proportion of the information in the original datasets that is explained by each new variable. Moreover, there are measures showing how important is each one of the original variables for describing the data.

In more detail the measures calculated are the following:

- The fit and loss values. The fit is calculated for the whole model and indicates how strong the correlation between the datasets is. The maximum fit ($fit_{\max}$) value equals the number of new dimensions (d) calculated by NLCCA and, if attained, indicates that the correlation of the data is perfect (i.e. a trivial case where a dataset is a transformation of the other). The loss measure on the other hand is calculated for each dimension and each dataset and shows how much of the variation of the original data can not be explained by the new variables (dimensions).
- The eigenvalue (E_i) for $i = 1,\ldots,d$ indicates how much of the correlation is explained by each dimension (i). It is calculated for each dimension and is equal to 1 minus the average loss for the dimension. The first new dimension of the NLCCA model has the maximum eigenvalue, the second the second highest and so on.
- The canonical correlation (ρ_i) is another measure that shows the strength of the correlation between the datasets for each new variable (dimension). It is usually used for two sets of variables, but it can easily be generalized for more than two. For two sets of variables, the canonical correlation per dimension is obtained by the following formula: $\rho_i = 2 \times E_i - 1$.
- The multiple fit ($mfit$) is used to show which of the original variables are more important for describing the information in the original datasets. Variables with high $mfit$ contribute better in describing the data. In our case, $mfit$ can help us choose the keywords and/or MeSH terms that better describe the information of the biomedical articles.

In order to better understand certain features of the data structure like trends, patterns, groupings or outliers we can plot in two dimensions the articles using the first new variables obtained from the NLCCA model. The graphical representations are based on scatterplots [13], a relatively simple method to represent combinations of variables. We can use them to simultaneously plot several pairs of variables and visualize their relations in a two dimensional space.

TABLE I
VALIDATION & INTERPRETATION MEASURES

Measure	Symbol
Total average loss	$loss_{avg}$
Average loss per dimension	$loss_{avgi}$
Maximum fit	$fit_{\max}$
Multiple fit	$mfit$
Model Dimensions (new variables)	d
Eigenvalue per dimension	E_i
Canonical Correlation per dimension	ρ_i

Based on earlier works [14], [15] we knew a priori that the selected articles for the experiment belonged to 12 groups, each corresponding to a specific Gene Ontology (GO) term (Table II) [16]. The GO terms constitute a structured vocabulary proper for describing gene product functions, like molecular functions, biological processes or cellular compartment. The scatterplots produced after the application of NLCCA revealed this grouping in great extend indicating that the new dimensions can discriminate efficiently the underlying GO groups.

TABLE II
GO TERMS RELATED TO THE DATASETS

GO no.	group	GO terms
1		Autophagy
2		Cell cycle
3		Cell proliferation
4		Cell cell signalling
5		Chemimechanical coupling
6		Meiosis
7		Metabolism
8		Oncogenesis
9		Stress response
10		Transport
11		Cell death
12		Signal transduction

To further investigate the grouping of the articles and visualize them, we used the centroids of the new variables obtained by NLCCA. These centroids are multidimensional points having as coordinates the means of the values of the new variables within a specific group.

III. EXPERIMENTATION

The experimentation involved the application of NLCCA to the datasets **A** and **B** produced by the data representation procedure, as described in the methodology section. Both datasets are based on 9009 biomedical articles. These articles have been used in our earlier work [15] for the development and evaluation of a classification model aiming to classify genes to 12 GO terms based on the information inside biomedical articles. We used the same articles in an attempt to further investigate the informative power of MeSH terms in relation to the article keywords.

The data representation procedure on the articles resulted in p=1642 keywords and therefore the first dataset (**A**) is essentially a 9009×1642 binary matrix with rows corresponding to each of the retrieved articles and columns to the keywords of the articles. Similarly, the second dataset (**B**) is a 9009×50 binary matrix with each column corresponding to one of the q=50 MeSH terms found to describe the whole body of data. Note that the MeSH terms used are all major topics of the 9009 articles and each one of them is relevant to at least 100 articles.

The results of NLCCA as applied to **A** and **B** datasets are shown in Table III (due to space limitation, results of dimensions 11 to 47 are not shown). The new variables (new dimensions) produced by the model are $d = 50$. Note that 50 is the maximum number of new variables that can be calculated for these datasets [10].

The fit of the model is 38.747 (close enough to 50) whereas the average loss over all dimensions is 11.253. Moreover, the eigenvalues vary from 0.927 to 0.704. We can also see from Table III that the first dimension explains most of the correlation (0.854) between the datasets. Furthermore, the last dimension describes a quite significant portion of the correlation (0.408). These results depict that the 50 new dimensions can explain much of the variation and information contained in both datasets.

TABLE III
SUMMARY OF NLCCA ANALYSIS

		Average Loss ($loss_{avgi}$)	Eigenvalue (E_i)	Correlation (ρ_i)	Fit
Dimension	1	0.073	0.927	0.854	
	2	0.106	0.894	0.788	
	3	0.121	0.879	0.758	
	4	0.128	0.872	0.744	
	5	0.137	0.863	0.726	
	6	0.139	0.861	0.722	
	7	0.149	0.851	0.702	
	8	0.157	0.843	0.686	
	9	0.159	0.841	0.682	
	10	0.169	0.831	0.662	
	…	…	…	…	
	48	0.292	0.708	0.416	
	49	0.295	0.705	0.41	
	50	0.296	0.704	0.408	
Sum		11.253			38.747

Average loss refers to the mean loss from dataset **A** and **B** .

The practical meaning of the analysis is that the 1642 binary variables corresponding to keywords and the 50 binary variables corresponding to MeSH terms of the articles can be efficiently replaced by only 50 new variables (or new dimensions) comprising a new dataset representing the biomedical articles.

Further analysis of the results involves descriptive statistics for the multiple fit measure of each keyword or MeSH term (Table IV& V).

TABLE IV
DESCRIPTIVE STATISTICS FOR *mfit*

Minimum	Maximum	Mean	Std. Deviation
0.009	1.497	0.0518	0.144

Table IV reports some basic statistics, such as minimum, maximum, mean and standard deviation of the fit, whereas Table V shows the 10 variables (keywords or MeSH terms) with the highest multiple fit values. Comparing the importance of keywords and MeSH terms, the latter have fit values above 0.5 whereas the keywords have values below 0.5. In other words, the MeSH terms seem to describe better the information contained in the biomedical articles.

The informative power of the first two new variables can be assessed from their discriminating ability, illustrated in the scatterplots of Figures 1 & 2. Figure 1 depicts the articles belonging to the 10^{th} and 11^{th} GO groups. Although there is some overlapping and some outlying points, the discrimination of the articles into two groups is clear. The zero value on the x-axis divides the two groups of articles. The categorization of the articles can also been seen from Figure 2. This figure shows the centroids of the articles for

the 12 GO groups in the space defined by the first two new variables.

TABLE V
THE TEN MOST IMPORTANT KEYWORDS & MESH TERMS

MeSH terms	*mfit*	Keywords	*mfit*
Drug effects	1.497	p53	0.48
Pharmacology	1.366	Oncogenes	0.425
Genetics	1.083	Drosophila	0.423
DNA Methylation	0.961	Apoptosis	0.419
Meiosis	0.914	Repair	0.418
Metabolism	0.899	Methylation	0.365
Physiology	0.891	Damage	0.329
DNA Repair	0.881	c.myc	0.291
Cell Transformation Neoplastic	0.879	Meiotic	0.284
Drosophila Proteins	0.870	Recombination	0.227

One can see that the centroids are positioned on the plot in a way that they can be distinguished. For example, the centroid of group 12 is quite far apart from the other groups, signifying that the articles belonging to group 12 has different information from the other articles. On the other hand, articles belonging to groups 9 and 6, for example, are not so different and their centroids can not be easily distinguished (at least in a 2-dimensional space).

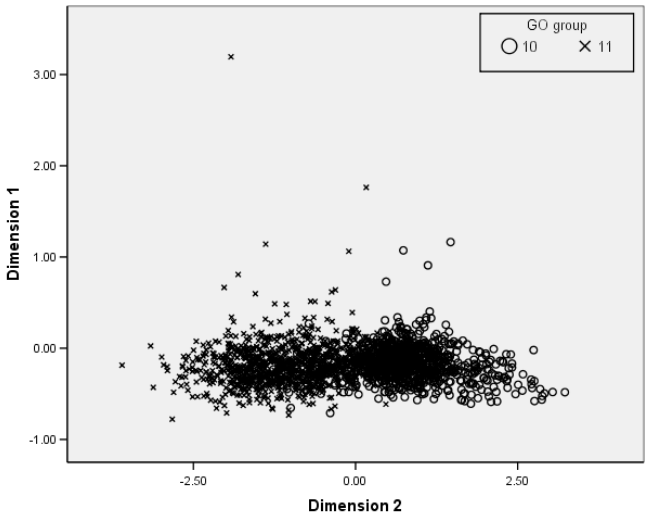

Fig. 1 The articles belonging to the 10th and 11th GO group plotted in a 2-dimensional space.

It must be noted that so far we only used the first two variables of the NLCCA model in order to produce the plots. In Figure 3, we can see that we can further use more of the new variables that are able to provide additional information about our datasets and help us distinguish better the articles.

As can been seen form Figure 3 the combination of the first, the second and the fourth new variables of the NLCCA model allows the better distinction between some of the centroids, compared to the scatterplot with only the first two variables. For example, the centroid of group 2 can be easily distinguished from the other groups when the 4th new variable is combined with the first two ones. Obviously, one should use as much as possible of the 50 new variables in order to perform a comprehensive study regarding the data structure exploration and interpretation. The important point however is that instead of using 1692 variables, the study becomes much simpler with the use of the 50 new variables.

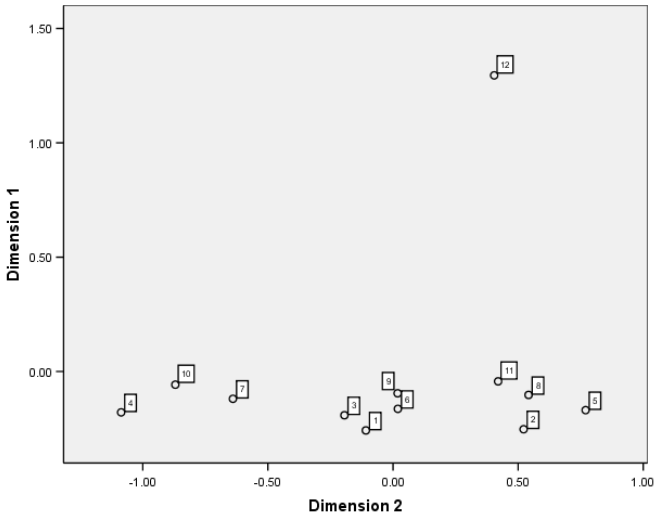

Fig. 2 Scatterplot of the centroids for the first two new variables (dimensions).

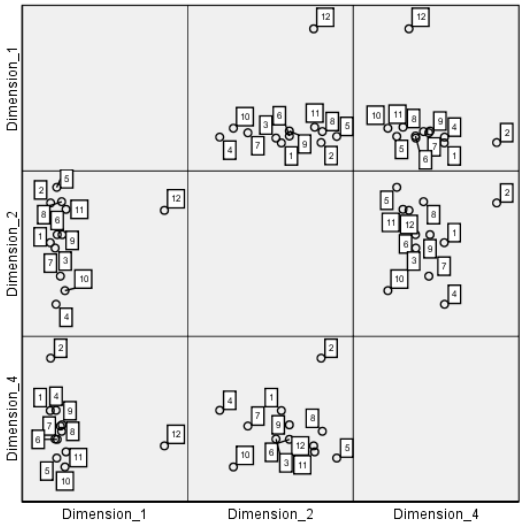

Fig. 3 A matrix of scatterplots for the 1st, 2nd and 4th new variable.

The fact that the articles can be grouped according to the GO groups they belong to, signifies that the articles contain specific keywords and MeSH terms which are closely related to the GO group they describe. After all, one can see from Table V that some of the most important variables of the model are related to some GO groups (Table II). For example, the "meiosis" MeSH variable is directly related to the "meiosis" GO term, whereas "DNA repair" is indirectly related to the GO terms "oncogenesis" and "cell proliferation". These examples show another interesting part of the proposed analysis, i.e. that several similar "semantic"

relations can be revealed from the statistical processing of the available data.

IV. Conclusions

It is evident from the experimentation that the non-linear canonical correlation analysis with optimal scaling can identify and model the correlations between the MeSH terms and the keywords contained in biomedical articles. It is based on a strong mathematical background and provides the tools to assess the correlations between different datasets. Furthermore, it allows us to explore the informative usefulness of the datasets by quantifying the importance of each variable for describing the articles. Another important advantage of the NLCCA method is that it can significantly reduce the number of variables contained in the two datasets, for example in our experiments the 1642 keywords and 50 MeSH terms are reduced to only 50 new variables. Also, NLCCA facilitates the visualization in two dimensions of multidimensional datasets in order to discover patterns and specific trends in graphical plots of the data.

The results from our experiments indicate that the new variables of the NLCCA model provide useful information about the data and help us clearly distinguish the articles based on the GO group they belong to. Another interesting result is that the single keywords have much less informative power for the articles than the MeSH terms.

It should be noted that the NLCCA with optimal scaling can be used as the first phase of a clustering or classification procedure such as SVM, naïve Bayes, discriminant analysis, etc [15]. The new variables can be further utilized in building classification models in order to categorize data. The interesting point here is that the new variables are uncorrelated and since many classification or clustering algorithms assume that the variables of the datasets are uncorrelated we can clearly see the advantage. Note also that in the case of keywords we have original variables highly correlated to another one, since each word depends on other words of the same text. Therefore, NLCCA can remove this correlation which might be a problem for building a model.

The fact that NLCAA can combine information from more than two different datasets makes it a very good method for biology, where there is usually the need to combine information from different data, like for example microarrays, biomedical articles, sequence and structure data.

As part of our future work we would like to apply NLCCA in the field of text clustering and more specifically in the field of biomedical article categorization. We would like to introduce a new methodology that could combine different sources of information in order to improve the efficiency and the usefulness of existing clustering methods, like K-means clustering. Another point that deserves further investigation is the application of NLCCA to real-valued vectors.

References

[1] B. H. Junker, C. Klukas, F. Schreiber, "VANTED: a system for advanced data analysis and visualization in the context of biological networks," BMC Bioinformatics. 2006 Mar 6;7:109.

[2] T. Yoneya, "PSE: a tool for browsing a large amount of MEDLINE/PubMed abstracts with gene names and common words as the keywords," BMC Bioinformatics. 2005 Dec 10;6:295.

[3] PubMed, National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institute of Health (NIH), http://www.pubmed.com

[4] C. A. Bean, R. Green, Ed. "Relationships in the organization of knowledge. NY: Kluwer Academic Publishers," 2001. pp. 171-184

[5] D.L. Rubin, C.F. Thorn, T.E. Klein, R.B. Altman. "A statistical approach to scanning the biomedical literature for pharmacogenetics knowledge," J Am Med Inform Assoc. 2005 Mar-Apr;12(2):121-9. Epub 2004 Nov 23. Erratum in: J Am Med Inform Assoc. 2005 May-Jun;12(3):364.

[6] M. Lee, W. Wang, H.Yu, "Exploring supervised and unsupervised methods to detect topics in Biomedical text," BMC Bioinformatics. 2006 Mar 16;7(1):140

[7] H.R. Strasberg, C.D. Manning, T.C. Rindfleisch, K.L.Melmon, "What's related? Generalizing approaches to related articles in medicine," Proc AMIA Symp. 2000;:838-42.

[8] P. Ruch, "Automatic assignment of biomedical categories: toward a generic approach," Bioinformatics. 2006 Mar 15;22(6) pp. 658-64.

[9] G. Salton, "Automatic text analysis", Science 168 (1970) pp. 335–343.

[10] G. Michailidis and J. de Leeuw. "The Gifi system for descriptive multivariate analysis," Statistical Science, 13 pp. 307--336, 1998.

[11] C. D. Manning, H. Schutze Foundations of Statistical Natural Language Processing, The MIT Press, 1999.

[12] J. J. Meulman, W. J. Heiser, SPSS Categories, SPSS Inc.

[13] J. F. Hair, R. E. Anderson, R. L. Tatham, W. C. Black, "Multivariate Data Analysis," 5th Edition, Prentice Hall PTR., 1998, ISBN: 0-13-930587-4.

[14] S. Raychaudhuri, J. T. Chang, P. D. Sutphin, R. B. Altman, "Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature," Genome Res 12 (1) (2002) pp. 203–214. URL http://dx.doi.org/10.1101/gr.199701

[15] T. Theodosiou, L. Angelis, A. Vakali, G. N. Thomopoulos, "Gene functional annotation by statistical analysis of biomedical articles," International Journal of Medical Informatics, to be published, URL http://dx.doi.org/10.1016/j.ijmedinf.2006.04.011

[16] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, M. Harris, D. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. Matese, J. Richardson, M. Ringwald, G. Rubin, G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," Nat Genet 25 (1) (2000n) pp. 25–29. URL http://dx.doi.org/10.1038/75556