BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–6

A web portal for multi-view text mining and vertical searches

Xinhai Liu [1,2,4]*, Olivier Gevaert [1,2,3], Léon-Charles Tranchevent [1,2], Yves Moreau [1,2], and Bart De Moor [1,2]

[1] Department of Electrical Engineering, ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, box 2446, 3001, Leuven, Belgium.
[2] IBBT-K.U.Leuven Future Health Department, Kasteelpark Arenberg 10, box 2446, 3001, Leuven, Belgium.
[3]Department of Radiology Center, Stanford University, USA.
[4] CISE& ERCMAMT,Wuhan University of Science and Technology, 430081, Wuhan, China

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Biomedical literature contains rich information that can be analyzed using different point of views. We propose a novel strategy to derive knowledge from textual information from a multi-view perspective. Our strategy has been applied to the MEDLINE corpus and analyzed using a disease based dataset. In particular, we investigated the effect of combining multiple views for clustering and assessed whether vertical searches can be more accurate for specific biological questions.

Results: Our results revealed that some redundancy can be observed between the different views. This phenomenon was expected since the views are sometimes not independent. However, they are also complementary and can be successfully used to significantly enhance clustering. In particular, integrating multiple controlled vocabularies or weighting schemes boosts the clustering performance in our case. A web application that implements our strategy has been developed. Multiple views can be used (different controlled vocabularies, weighting schemes, biomedical subjects) to characterize the gene set input by the user. Alternatively, vertical searches can be performed to answer specific biological problems by using the dedicated views that will restrict the search space. The output consists of gene-by-term profiles that can be displayed either as term clouds or as similarity matrices (color map). In addition, the hierarchical clustering of the input genes is performed. Results can be downloaded for further use.

Availability: The web application of our strategy is available: http://aulne8.esat.kuleuven.be/TextPrior/

Contact: xinhai.liu@esat.kuleuven.be

1 INTRODUCTION

1.1 The importance of text mining in biomedical world

Text mining helps biologist to collect structured biomedical knowledge automatically from large volumes of biological literature. During the past ten years, there was a surge of interest in automatic exploration of the biomedical literature, ranging from the modest approach of annotating and extracting keywords from text [14] to more ambitious attempts such as Natural Language Processing [4], and text mining based network construction and inference [15]. One of the main objectives of text mining is to structure the knowledge contained in the biological literature in order to extract biological entities and relations between them. In particular, these efforts effectively help biologists to identify the most likely disease candidate genes for further experimental validation [23]. It is often the case that text mining data is combined with other biological data within an elaborated workflow. For instance, text mining can serve as prior information for typical clinical decisions support algorithms such as Bayesian networks [1]. It is also possible to unify heterogeneous data sources such as clinical data with text mining based data sources [8].

1.2 Multi-view text mining

In general, a successful text mining approach relies much on an appropriate mining model, and the efficiency of biomedical knowledge discovery varies greatly between different models. Which text mining model is superior depends on the problem under consideration. This makes multi-view models more suited since they are more flexible to answer various biological applications. In our early work, we propose a multi-view text mining model based on the use of several controlled vocabularies [23]. We now propose to also consider the use of several term scoring (weighting) schemes, and the mining of distinct document corpus as additional views. More precisely, we define distinct document corpus by distributing the journals based on their biomedical subjects, or by grouping the papers based on their publication year. The different views are redundant but also complementary. Therefore the integration of multiple views is expected to allow for a more accurate definition of our current knowledge in genetics and medicine. Another motivation behind our work is to provide a vertical search engine, in order to get insight into specific biomedical fields [2]. In contrast to general search engines that attempt to index large portions of the World Wide Web or whole

*to whom correspondence should be addressed

© Oxford University Press 2005.

databases, vertical search engines typically attempt to index only the documents that are relevant to a pre-defined topic. In our case, this segment can be defined by selecting one or several biomedical subjects, vocabularies, or time periods.

1.3 Related work

The concept of multi-view document analysis was originally proposed by Bickel and Scheffer who describe a web document clustering strategy that combines intrinsic view of web pages (text based similarity) and extrinsic view (citation link based similarity) [3]. More recently, Gaulton *et al.* have adopted three different ontologies on eight text sources and built the CAESAR system that annotates human disease genes and identifies potentially novel disease genes [7]. Lately, Névéol *et al.* have combined three different models (dictionary lookup, post-processing rules and NLP rules) to identify MeSH main heading/subheading pairs from medical text [19]. Much effort has been put into the automatic extraction of disease gene relations from free text [18, 5]. To improve the performance of mapping biomedical sentences into an ontology, Kim *et al.* proposed an integrated information retrieval technique that combines a simple language model with document frequencies and a distance measure, and followed by clustering [13]. In 2005, we implemented a framework called TXTGate that combines literature indices of selected public biological resources in a flexible text-mining system designed towards the analysis of gene sets [8]. More recently, we have used multi-view text mining data for gene prioritization and clustering [23]. Our work shares the same flavor, however we extend multi-view to broad concepts and emphasize vertical search from certain specific perspective. When compared to our former multi-view text mining work, our research brings three novel items: extension of the multi-view concept to a broad and flexible framework, implementation of vertical text mining from multiple perspectives, and a tensor based data fusion method. In the current study, we extend the multi-view concept to the use of several weighting schemes in addition to the use of several vocabularies. We also propose a vertical search engine by restricting the text mining analysis to a subset of the original document corpus. The subset can be defined either by biomedical subjects or by publication time periods, and only the relevant papers are then indexed. We have implemented this scheme into a freely available computational framework that can be used to investigate genes or gene sets through similarity analysis and clustering.

2 MATERIALS AND METHODS

2.1 Document corpus

One of the most important resources for biological text mining applications is the National Library of Medicine's bibliographic database (MEDLINE). MEDLINE contains more than 18 millions publications that cover many aspects of biology, chemistry and medicine. There is almost no limit to the types of information that may be recovered through careful and exhaustive mining. There are more than 10,000 biomedical related journals, accumulating over 700,000 new publications each year. In the current study, we use the MEDLINE repository as of April 2010. Each publication is represented by its title and its abstract (when available). The full article is never retrieved. The mapping between genes and publications from Entrez GeneRIF was used to index the MEDLINE repository. The GeneRIF data was also collected in April 2010, and consists of 290,000 associations between 13,633 human genes and 322,639 MEDLINE publications (from 3,276 journals).

2.2 Indexing

In the first step, documents are indexed and a document-by-term matrix is computed. The indexing process is performed using the Java Lucene package [9], and more details can be found in our earlier work [23]. In the second step, we averagely combine the document-by-term vectors to obtain gene-by-term vectors according to the GeneRIF mapping. Each feature of the gene vector then corresponds to the score of a term from a fixed vocabulary (ontology). The multiple views adopted in this research refer to different weighting schemes, controlled vocabularies, and biomedical subjects.

2.2.1 Weighting schemes A weight is a statistical measure used to evaluate how important a term is to a document in a corpus. The importance increases proportionally to the number of times this term appears in the document but is offset by its frequency within the whole corpus. In the current study, we used three different weighting schemes: TF (Term Frequency), IDF (Inverse Document Frequency), and TFIDF (Term Frequency * Inverse Document Frequency). More details about these weighting schemes can be found in **Supplementary material 3**. TFIDF is often used in information retrieval and text mining, but IDF is also found to work well in biomedical related text mining [23]. Since it is hard to estimate beforehand which scheme is universally superior, both are made available. In addition, TF is also proposed but mainly for comparative studies since it was shown to give less meaningful results [17].

2.2.2 Controlled vocabularies We have selected four vocabularies from four bio-ontologies. Three of them (GO, MeSH, OMIM) have proved their merit in our earlier work [24]. In addition, we have also selected an ontology from the National Cancer Institute (NCI) to cover more specifically cancerous diseases. These four ontologies are briefly introduced in **Supplementary material 2**. The ontological terms are first extracted, stored as bag-of-words, and then preprocessed for text mining. This pre-processing includes transformation to lower case, segmentation of long phrases, and stemming. After preprocessing, these vocabularies are fed into a Java program based on the Apache Java Lucene API to index the titles and abstracts of MEDLINE publications relevant to human genes.

2.2.3 Biomedical subjects The National Library of Medicine (NLM) assigns MeSH terms to each journals to describe their main focus. Not all journals are associated to MeSH terms, and we therefore only keep the journals with at least one term (52 journals discarded over 3,276 journals in total). There are in total, 114 distinct MeSH terms used to define the journal's scope, and there are sometimes several terms per journal. The distribution of the 114 MeSH terms is heavily biased. For instance, the term with the largest number of publications is 'Molecular biology', with 33,164 publications. At the other end of the spectrum, 'Optometry' is only linked to a single paper. More details about these MeSH terms and

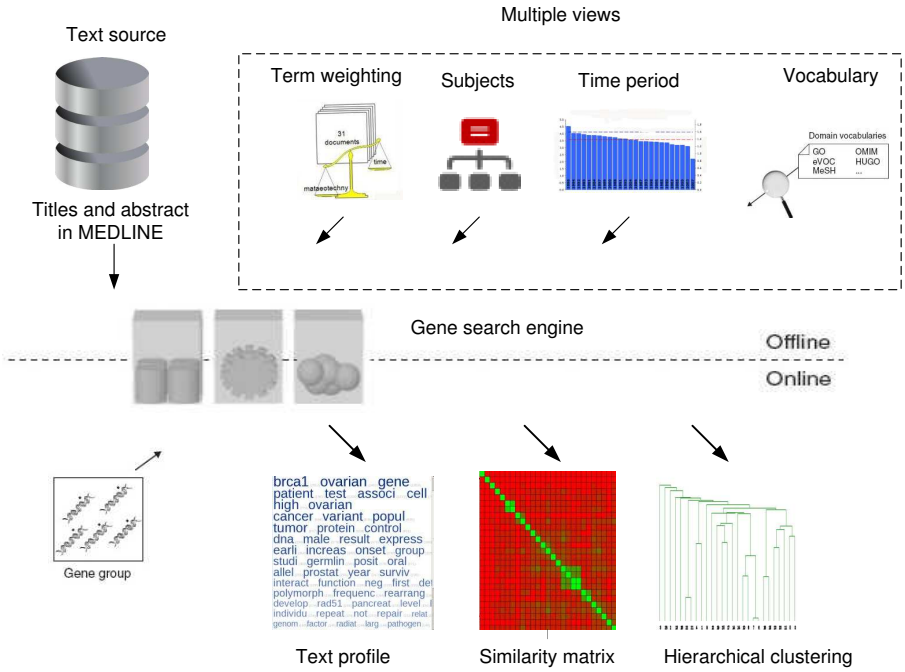

Fig. 1. Conceptual overview of our text mining system. The whole corpus is indexed with several vocabularies, weighting schemes, biomedical subjects and publication time periods (multiple views). Sets of genes can then be investigated on-line: the text profiles of the genes are retrieved. Furthermore, similarity matrices can be computed and hierarchical clustering is performed.

the associated number of publications can be found in Table 4 of **Supplementary material 5**.

2.2.4 Publication year For the current MEDLINE, the publication year ranges from 1950 to 2010. Notice that 5,537 papers have been removed since their publication year is missing. The yearly paper distribution (human gene related) is shown in Figure 1 of **Supplementary material 1**. As expected, the number of papers that are linked to human genes is increasing since the sequencing of the human genome. We have roughly divided the papers into four categories according to the publication year: 1950-1990, 1980-2000, 2001-2005, and 2006-2010 (see also Table 3 of **Supplementary material 4**.).

2.3 Web application

The Web application was developed using the Google Web Toolkit Version 2.0[2]. A conceptual overview of our system is illustrated in Figure 1. It can be fed with a set of genes and returns the text profiles of these genes as well as the similarity matrix and the associated clustering results. These results can be downloaded for further analysis. For each query gene of the input gene set, a text profile is retrieved. This profile contains the annotation terms and the corresponding scores. It is possible to display the top 10 terms that are annotated to the genes (terms with the highest scores). It is also possible to visualize the profile as a tag cloud (or term cloud), for which the font size of a term is proportional to its score. To compare gene profiles, we compute the cosine similarity between the two corresponding gene-by-term vectors. We offer the possibility to cluster on-line the gene set by means of hierarchical clustering. The clustering is performed in Java (own implementation) using the average link [12] and the aforementioned similarity measure. The hierarchical structure can also be visualized to allow for an exploratory clustering strategy. For convenience, the clustering can only be achieved with 100 genes or less.

[2] http://code.google.com/webtoolkit/

2.4 Hybrid clustering approach

Clustering is helpful to identify the functional relationship between genes [10]. In this study, we apply a clustering strategy in order to assess whether combining multi-views leads to an increased performance. Hybrid clustering refers to joint clustering that integrates multi-view data, and is expected to boost the clustering performance. The hybrid clustering strategy we adopted is based on a tensor method of higher-order singular value decomposition and is therefore named HC-HOSVD [16]. In this research, we employ the normalized similarity matrix for the formulation of HC-HOSVD, instead of modularity matrix in the original formulation. Suppose the relaxed cluster indicator matrix $\mathbf{U} \in \mathbb{R}^{N \times M}$ where N is the number of data points and M is the number of clusters. Since the normalized similarity matrix $\mathbf{S}_N$ is positive semidefinite, the spectral clustering can be re-formulated as a Frobenius-norm optimization problem,

$$\begin{aligned} &\max_{\mathbf{U}} \|\mathbf{U}^T \mathbf{S}_N \mathbf{U}\|_F^2, \\ &\text{s.t. } \mathbf{U}^T \mathbf{U} = \mathbf{I}. \end{aligned} \tag{1}$$

where the Frobenius norm, $\| \mathbf{A} \|_F$, of a matrix $\mathbf{A} = (\mathbf{A})_{ij}$ is given by $\| \mathbf{A} \|_F = \sqrt{\sum_{i,j} \mathbf{A}_{ij}^2}$. For each single-view data, we can formulate the spectral clustering optimization as (1) to get the corresponding partition. In fact, the multi-view data can be utilized together to find a joint partition, which is assumed to be beyond each individual partition. This is the basic concept underlying hybrid clustering. A natural way is to integrate each optimization of spectral clustering. Thus, we formulate the optimization of hybrid clustering by linearly combining each individual optimization as

$$\max_{\mathbf{U}} \sum_{i=1}^{K} \|\mathbf{U}^T \mathbf{S}_N^{(i)} \mathbf{U}\|_F^2, \quad \text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I}. \tag{2}$$

where $\mathbf{S}_N^{(i)}$ is the normalized similarity matrix of single-view data and U is the common factor shared by multi-view data, that can be partitioned by the common clustering methods, such as *k*-means, to obtain the final clustering labels. This optimization can be approximated by higher-order Singular Value Decomposition (HOSVD)[6] that can be seen as the multi-view version of principal component analysis (PCA). More details can be found in **Supplementary material 6**.

2.5 Validation data

We validate our approach with the human disease benchmark data set of Endeavour [22], from which we selected 14 diseases and the 264 associated genes. The 14 diseases are presented in Table **??**. To compare different views, the cosine similarities between all gene-by-term vectors are computed for each views, which leads to the generation of one similarity matrix per view. The cosine similarity is then computed between these two matrices and used as an estimate of the global similarity of the two underlying views [21]. Given two similarity matrices $\mathbf{S}_i$ and $\mathbf{S}_j$, and their corresponding vectorizations are vec($\mathbf{S}_i$) and vec($\mathbf{S}_j$), where vec($\mathbf{S}$) means all columns of $\mathbf{S}$ are stacked each other, the cosine similarity cross two views is computed as,

$$\cos(\theta_{i,j}) = \text{vec}(\mathbf{S}_i) \times \text{vec}(\mathbf{S}_j) / (\|\text{vec}(\mathbf{S}_i)\|_2 \times \|\text{vec}(\mathbf{S}_j)\|_2) \tag{3}$$

where $\cos(\theta_{i,j}) = \cos(\theta_{j,i})$ and it ranges from 0 (two views are completely different) and 1 (two views are identical).

Regarding clustering evaluation, the gene data sets used in our experiments are provided with disease labels, therefore the clustering performance is evaluated by comparing the automatic partitions with the labels using Adaptive Rand Index (ARI [11]) and Normalized Mutual Information (NMI [20]). We set the cluster number K to 14 since there are 14 diseases.

3 RESULTS

This section presents a similarity analysis performed on the individual views, a benchmark of the method based on clustering, and introduces our web tool.

3.1 The similarities among multiple views

Each view provides text information from a certain perspective. In this section, before combining multiple views, we measure the similarities or the differences among these views.

Vocabulary	GO	MeSH	OMIM	NCI
GO	1	0.7925	0.8499	0.8966
MeSH	0.7925	1	0.7565	0.8111
OMIM	0.8499	*0.7565*	1	0.8192
NCI	**0.8966**	0.8111	0.8192	1

Table 1. The cosine similarity between the four vocabularies. The largest non-self similarity is shown in bold, the smallest non-self similarity is shown in italics.

Weighting scheme	TF	IDF	TFIDF
TF	1	0.7326	0.8715
IDF	*0.7326*	1	0.8524
TFIDF	**0.8715**	0.8524	1

Table 2. The cosine similarity between the three weighting schemes. The largest non-self similarity is shown in bold, the smallest non self similarity is shown in italics.

3.1.1 Similarities among vocabularies To compare the different vocabularies, we set the weighting scheme to IDF, and the whole corpus was indexed. The global similarities among the four vocabularies are shown in Table 1. The largest similarity exists between GO and NCI (0.8966) while the smallest similarity between MeSH and OMIM (0.7565). Altogether, the results indicate that although there are differences among the vocabularies, these are not huge. Similar results are obtained with TFIDF (data not shown).

3.1.2 Similarities among weighting schemes To analyze multiple weighting schemes, we used MeSH as the vocabulary, and the whole corpus was indexed. The global similarities among the three weighting schemes are presented in Table 2. The largest similarity exists between TF and TFIDF (0.8715), the smallest similarity between TF and IDF (0.7326). Similar results are obtained with different vocabularies (data not shown).

3.1.3 Similarities between biomedical subjects In order to compare the different biomedical subjects, we used MeSH as the vocabulary, we set the weighting scheme to IDF, and we selected all publication time periods. Among the 114 biomedical subjects, we selected the six that are associated with the largest number of papers. The global similarities among these six subjects are shown in Table 3. The largest similarity exists between Molecular Biology and Cell biology (0.7919), the smallest similarity between Allergy & Immunology and Genetic medical (0.2212). As can be observed from the Table 3, the use of different subjects gives more different results than the use of different vocabularies or weighting schemes. It also motivates vertical searches that are able to provide precise and unique information.

3.1.4 Influence of the biological question We have also performed the analysis on a set of diabetes related genes to investigate the differences among the multiples views. The details of these multi-view text mining profiles can be found in **Supplementary material 8** (Supplementary Tables 7, 8, 9, 10 and 11). We can see that, for diabetes, there is less overlap among the multiple views in general,

Medical subject	Allergy & Immunology	Cell biology	Genetic medical	Molecular biology	Neroplasms	Science
Allergy & Immunology	1	0.3527	0.2212	0.3032	0.3958	0.3657
Cell biology	0.3527	1	0.5080	0.7919	0.7471	0.5574
Genetic medical	*0.2212*	0.5080	1	0.7766	0.4634	0.4967
Molecular Biology	0.3032	**0.7919**	0.7766	1	0.6900	0.5843
Neroplasms	0.3958	0.7471	0.4634	0.6900	1	0.5613
Science	0.3657	0.5574	0.4967	0.5843	0.5613	1

Table 3. The cosine similarity of multi-view subjects. Except the self-similarity, the largest similarity is shown in bold while the smallest similarity is shown in italics.

Vocabularies	NMI	P-value	ARI	P-value
Combined	**0.7290± 0.02**	—-	**0.5393±0.05**	—-
GO	0.5537± 0.01	4.4e-39	0.3575±0.03	1.34e-23
MeSH	0.7012±0.02	1.14e-7	0.5157±0.05	0.0316
OMIM	0.6893±0.02	7.01e-12	0.4868±0.05	8.42e-6
NCI	0.5109±0.01	2.97e-46	0.2844±0.02	3.10e-34

Table 4. Clustering results for multiple vocabularies. The mean values and standard deviations are observed from 50 repetitions. The best values are shown in bold. 'Combined' refers to the hybrid clustering of the integration of MeSH and OMIM.

Weighting scheme	NMI	P-value	ARI	P-value
Combined	**0.7001±0.02**	——	**0.5236±0.05**	————
TFIDF	0.6868±0.02	0.0017	0.5021±0.04	0.0466
TF	0.4963±0.02	4.52e-49	0.2882±0.02	1.14e-35
IDF	0.6872±0.01	2.85e-4	0.5039±0.04	0.0232

Table 5. Clustering results for multiple weighting schemes. The mean values and standard deviations are observed from 50 repetitions. The best values are shown in bold. 'Combined' refers to the hybrid clustering of the integration of TFIDF and IDF data.

and in particular for vocabularies and biomedical subjects. This analysis indicates that different results can be obtained with different biological questions.

3.2 Multi-views clustering

In this Section, we compare the clustering results when applied on a single view and when applied on multiple views.

3.2.1 Clustering using multiple vocabularies Using the four different vocabularies, we build four different gene-by-term matrices. Therefore, four normalized similarity matrices are generated in total. We then applied our hybrid clustering method (HC-HOSVD) to combine the multiple views and compare to the use of any single view. The results are presented in Table 4. It can be observed that the best single view performance is obtained by using MeSH (NMI 0.7012, ARI 0.5157). However, the integration of multiple views, MeSH and OMIM in this case, (NMI 0.7290, ARI 0.5393) is significantly superior to the use of MeSH or OMIM alone. These results demonstrate that the integration of multiple vocabularies is able to enhance the clustering performance. Similar results are obtained with TFIDF weighting scheme (data not shown).

3.2.2 Clustering using multiple weighting schemes In this paper, we have implemented three weighting schemes: TF, IDF and TFIDF, which allows us to get three gene-by-term matrices. We expect that integrating this type of multiple views will enhance the clustering performance. The results are presented in Table 5. The best performance for single view is obtained by TFIDF (NMI 0.7001, ARI 0.5236), just slightly ahead of IDF (NMI 0.6872, ARI 0.5039). However, the hybrid clustering using multiple views is still significantly superior, showing that the integration of multiple weighting schemes , TFIDF and IDF in our case, can boost the clustering performance as well. Similar results are obtained with other vocabularies (data not shown).

	GO (37,069)	MesH (29,709)	OMIM (5,021)	NCI (27,247)
GO (37,069)	—	9,952	1,431	5,409
MesH (29,709)	9,952	—	3,191	6,399
OMIM (5, 021)	1,431	3,191	–	1,071
NCI (27,347)	5,409	6,399	1,071	–

Table 6. Term overlaps between the four vocabularies. The total number of terms for each vocabulary is denoted between brackets.

4 DISCUSSION

We have developed a literature based gene retrieval system that is able to provide multi-view observations as well as vertical search. The aim of our system is to aid the clinical analysis and biomedical research. We illustrate its usefulness through a clustering validation that proved the efficiency of the multi-view strategy. With respect to vertical search, our search engine is able to help the users who want very specific knowledge (corresponding to one of the several branches of the whole biomedical world). Biomedical research is a fast developing field, it is therefore divided into more and more tiny specialized fields, hence, such a system that proposes precise searches is really more and more required.

Based on the similarity analysis of multi-view text mining, as can be seen in Table 1 (multi-view vocabularies), Table 2 (multi-view weighting schemes) and Table 3 (multi-view biomedical subjects), the views appear different but redundant. This redundancy was expected because the multiple vocabularies we used share common terms as denoted in Table 6. The largest overlap is observed between MesH and OMIM, with 3,191 common terms, which represents 64% of OMIM. In addition, the multiple weighting schemes we used also share part of their formulas as presented in **Supplementary material 3**. However, beside this redundancy, we can observe that the integration of multiple views is almost always leading to a better representation of the data, therefore leading to better validation results.

Regarding clustering of multi-view data, tensor based hybrid clustering method is able to make the best of the data. As long as the multi-view data has complementray information and that the noise level is kept under control, the combination with the tensor based strategy is always able to improve the clustering performance.

The idea of multi-view text mining is not restricted to the several views mentioned in this study. For example, according to the citation impact factor of the related journals and their citations, the papers in MEDLINE could be classified into different categories, which thus would correspond to other types of views for text mining. One research avenue to explore in the future, we can use gene-by-concept vector by latent semantic analysis [10], which is expected to identify the implicit relationship among genes.

5 CONCLUSION

In this paper, we have developed a Web based system that can be used to profile a gene set from a text-mining point of view. On the one hand, the information from multiple views can be combined to provide rich and complementary information. On the other hand, information from a specific view offers a vertical observation with a specific focus. The system can be utilized to identify the relationships between genes to aid the clinic diagnosis straightforwardly or to provide text prior information for further analysis. We have benchmarked the overall approaches with a set of disease genes. The results demonstrate the power of combining multiple views when performing clustering due to the synergic effect of the fusion. However, we also observed that better results can be obtained for a specific biological question when using a single highly relevant view. In the further research, we plan to apply our system to enhance candidate gene prioritization. Meanwhile, we are also planning an extension of the approach to other biomedical entities, such as diseases or biological pathways.

ACKNOWLEDGEMENT

Research supported by (1) China Scholarship Council (CSC, No. 2006153005); (2)Research Council KUL: GOA Ambiorics, GOA MaNet, CoE EF/05/006 Optimization in Engineering(OPTEC), IOF-SCORES4CHEM, several PhD/postdoc & fellow grants; (3)FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (Tensors), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC); (4)IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare; (5)Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011); (6)EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940); (7)Contract Research: AMINAL; Other: Helmholtz: viCERP; ACCM; Bauknecht; Hoerbiger;

REFERENCES

[1]P. Antal, G. Fannes, D. Timmerman, Y. Moreau, and B. De Moor. Using literature and data to learn bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in Medicine*, 30(3):257–281, 2004.

[2]J. Battelle. *The Search: How Google and its Rivals Rewrote the Rules of Business and Transformed Our Culture*. Portfolio, New York, 2005.

[3]S. Bickel and T. Scheffer. Multi-view clustering. In *Proceedings of the Fourth IEEE International Conference on Data Mining*, ICDM '04, pages 19–26, Washington, DC, USA, 2004. IEEE Computer Society.

[4]J. Björne, F. Ginter, S. Pyysalo, J. Tsujii, and T. Salakoski. Complex event extraction at PubMed scale. *Bioinformatics (Oxford, England)*, 26(12):382–390, 2010.

[5]H. Chun, Y. Tsuruoka, J. Kim, R. Shiba, N. Nagata, and T. Hishiki. Extraction of gene-disease relations from medline using domain dictionaries and machine learning. In *In Proceedings of the 11th Pacific Symposium on Biocomputing*, pages 4–15, 2006.

[6]L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4):1253–1278, 2000.

[7]K. J. Gaulton, K. L. Mohlke, and T. J. Vision. A computational system to select candidate genes for complex human traits. *Bioinformatics*, 23:1132–1140, 2007.

[8]P. Glenisson, B. Coessens, S. Van Vooren, J. Mathys, Y. Moreau, and B. De Moor. Txtgate: profiling gene groups with text-based information. *Genome Biology*, 5(6):R43, 2005.

[9]E. Hatcher and O. Gospodnetić. *Lucene in Action*. Manning Publications Co., 2004.

[10]R. Homayouni, K. Heinrich, L. Wei, and M. W. Berry. Gene clustering by latent semantic indexing of medline abstracts. *Bioinformatics*, 21(1):104–115, 2005.

[11]L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[12]A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[13]M. Kim, Q. Dou, O. R. Zaiane, and R. Goebel. Unsupervised mapping of sentences to biomedical concepts based on integrated information retrieval model. In *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, pages 322–329, New York, NY, USA, 2010. ACM.

[14]M. Krallinger, R. Alonso-Allende, and A. Valencia. Text-mining approaches in molecular biology and biomedicine. *Drug Discovery Today*, 10:439–445, 2005.

[15]S. M. Leach, H. Tipney, W. Feng, W. A. Baumgartner, P. Kasliwal, R. P. Schuyler, T. Williams, R. A. Spritz, and L. Hunter. Biomedical discovery acceleration, with applications to craniofacial development. *PLoS computational biology*, 5(3):e1000215+, 2009.

[16]X. Liu, L. De Lathauwer, F. Janssens, and B. De Moor. Hybrid clustering on multiple information sources via hosvd. In *Proceedings of the 7th International Symposium on Neural Networks (ISNN 2010)*, pages 337–345, 2010.

[17]X. Liu, S. Yu, F. Janssens, W. Glänzel, Y. Moreau, and B. De Moor. Weighted hybrid clustering by combining text mining and bibliometrics on a large-scale journal database. *Journal of the American Society for Information Science and Technology*, 61(6):1105–1119, 2010.

[18]A. Névéol, W. Kim, W. J. Wilbur, and Z. Lu. Exploring two biomedical text genres for disease recognition. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 144–152, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[19]A Névéol, SE Shooshan, SM Humphrey, TC Rindflesh, and AR. Aronson. Multiple approaches to fine-grained indexing of the biomedical literature. In *Pac Symp Biocomput*, pages 292–303. World Scientific, 2007.

[20]A. Strehl and J. Cluster Ghosh. Ensembles-a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3, 2002.

[21]P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005.

[22]S. Yu, X. Liu, L. C. Tranchevent, W. Glänzel, J. Suykens, B. De Moor, and Y. Moreau. Optimized data fusion for k-means laplacian clustering. *Bioinformatics*, 27(1):118–126, 2010.

[23]S. Yu, L. Tranchevent, B. De Moor, and Y. Moreau. Gene prioritization and clustering by multi-view text mining. *BMC bioinformatics*, 11:28, 2010.

[24]S. Yu, S. Van Vooren, L. Tranchevent, B. De Moor, and Y. Moreau. Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining. *Bioinformatics*, 24(16):i119–i125, 2008.