On the potential of domain literature for clustering and Bayesian network learning

Peter Antal*
Katholieke Universiteit Leuven
El. Eng. ESAT-SCD (SISTA)
Kasteelpark Arenberg 10
B-3001 Leuven, Belgium
peter.antal@esat.kuleuven.ac.be

Patrick Glenisson*
Katholieke Universiteit Leuven
El. Eng. ESAT-SCD (SISTA)
Kasteelpark Arenberg 10
B-3001 Leuven, Belgium
patrick.glenisson@esat.kuleuven.ac.be

Geert Fannes
Katholieke Universiteit Leuven
El. Eng. ESAT-SCD (SISTA)
Kasteelpark Arenberg 10
B-3001 Leuven, Belgium
geert.fannes@esat.kuleuven.ac.be

ABSTRACT

Thanks to its increasing availability, electronic literature can now be a major source of information when developing complex statistical models where data is scarce or contains much noise. This raises the question of how to integrate information from domain literature with statistical data. Because quantifying similarities or dependencies between variables is a basic building block in knowledge discovery, we consider here the following question. Which vector representations of text and which statistical scores of similarity or dependency support best the use of literature in statistical models? For the text source, we assume to have annotations for the domain variables as short free-text descriptions and optionally to have a large literature repository from which we can further expand the annotations. For evaluation, we contrast the variable similarities or dependencies obtained from text using different annotation sources and vector representations with those obtained from measurement data or expert assessments. Specifically, we consider two learning problems: clustering and Bayesian network learning. Firstly, we report performance (against an expert reference) for clustering yeast genes from textual annotations. Secondly, we assess the agreement between text-based and data-based scores of variable dependencies when learning Bayesian network substructures for the task of modeling the joint distribution of clinical measurements of ovarian tumors.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning—knowledge acquisition; I.5.3 [**Pattern Recognition**]: Clustering

Keywords

Text mining, data mining, clustering, Bayesian networks

*These authors contributed equally to this work.

SIGKDD '02 Edmonton, Alberta, Canada

1. INTRODUCTION

In many complex knowledge discovery problems, such as identifying relationships between a large number of genes in genomics or between clinical measurements in medical informatics, knowledge about the domain variables and relationships between these variables is fragmentary at best, cost of data collection is high, and measurements are often noisy or unreliable. When setting up such models, domain literature is invaluable as it often contains a lot of information, albeit fragmentary, about the problem at hand. Further, electronic literature is easy to process, although extracting information from it still is a major challenge.

In this paper we reach one step further than classical text mining and attempt to integrate textual information into the modeling process on an equal footing with statistical data. We investigate whether similarities or dependencies between variables quantified from textual information represented by shallow statistic vectors agree with those identified by expert assessment or measurement data. In particular, we characterize which text representations (boolean, frequency, or term frequency–inverse document frequency) and statistical scores of variable similarity or dependency best support the use of literature in clustering and Bayesian network learning.

As a first case, we cluster a custom collection of yeast genes from textual annotations extracted from databases of gene information (and possibly expanded with literature abstracts). We perform clustering using the k-medoids algorithm with a similarity measure derived from the cosine similarity. We assess the agreement between the resulting clusters and an expert reference using the adjusted *Rand* index. As a second case, we consider the task of modeling the dependencies between clinical measurements of ovarian tumors and learn Bayesian network substructures using expert annotations (possibly expanded with literature abstracts). We introduce a new text-based score of local dependency. We assess the agreement between text-based scores of local dependency and data-based scores and an expert assessment using correlation coefficients and Spearman rank correlation coefficients.

As a conclusion, we observe in both cases that the information extracted from textual sources captures an important part of the information present in the data or provided by the expert. We also conclude that different sources of annotations and different text representations have widely varying performance (which is also problem specific). Thus,

finding the most effective textual source of information and the best text representation is essential if we want to integrate text and data in knowledge discovery.

The paper is organized as follows: Section 2 presents a framework for the integrated analysis of data, domain knowledge, and literature with an emphasis on the evaluation of the use of literature in statistical methods. Section 3 summarizes the text representations, relevance measures, and general linguistic preprocessing used in the paper. Section 4 discusses the usage of literature in clustering expression data and overviews the genomic information sources for the model organism yeast. Quantitative measures for the usability of literature in clustering are introduced. Section 5 presents the medical problem of assessing ovarian tumors by ultrasonography together with the task of identifying a probabilistic model of the corresponding clinical measurements. We introduce Bayesian networks together with a standard score based on data for the identification of such models. We also introduce a new score based on literature that plays a role similar to the previous score in identifying Bayesian network substructures but this time from literature instead of data. Section 6 presents the comparison of literature-based clustering against expert knowledge in the yeast genomic domain and the comparison of the literature- and data-based scores used in learning Bayesian networks in the ovarian cancer domain. Results are reported for several vector representations of text and types of textual information sources. Specifically, we report results on the use of automatically expanding the initial annotation of the variables. Finally, Sections 7 and 8 contains the discussion, conclusion, and our view on how the integrated use of literature and statistical data is possible.

2. A FRAMEWORK FOR THE ANALYSIS OF TEXT, DATA, AND PRIOR

In most domains the information that can be used in modeling comes from different types of sources. On the one hand, we have the observed cases, which lead to a data set $\underline{D}$. This type of information is the most straightforward to work with. On the other hand, a lot of prior domain knowledge can be available in various formats. In this paper we will restrict this prior knowledge to (1) textual information and (2) a small amount of expert knowledge for validation. Textual information is hard to deal with in computational and statistical procedures and it needs a lot of preprocessing to convert into a usable format, but it can be valuable when only a few data samples are present or the data is noisy. The optimal but most difficult strategy would be to use both information sources in the model building process in some integrated fashion. To evaluate the possibilities of such combined methodologies, we investigate the agreement between data-based scores commonly used in Bayesian network model selection and a newly introduced text-based score explained in Section 5. Additionally, we compare the results of a text-based clustering with a gold standard provided by an expert, as explained in Section 4.

In both cases we have a set of domain variables $V_1, \ldots, V_m$, which represent medical observations in the Bayesian network case and yeast genes in the cluster case. For these variables we want to derive somehow a *relatedness measure* based on textual information. To achieve this, an expert annotates these domain variables with free text (describing the variables) and relevant references for this variable in the literature repository. The following step converts these textual annotations into a vector representation used in the experiments explained above. It is expected that there will be no strict match between the textual information and the data or prior knowledge, but to some extent they should reveal the same relations. In the presented framework, we hope to demonstrate that both information sources can complement each other in an integrated model building process.

3. CONCEPTS FROM INFORMATION RETRIEVAL

We assume that we have a free-text annotation for each domain variables. Such an annotation can further contain references to domain-related document collections. These two types of annotations give rise to the commonly used vectorial text representations and the reference representation.

3.1 Representations of annotations

The representation called the vector space model encodes a document in a k-dimensional space where each component represents a corresponding word, neglecting the grammatical structure of the text. We applied the Porter stemmer to canonize the words [6], the synonyms were replaced, and the phrases were detected and merged. In both domains, we automatically constructed a large vocabulary containing more than one million words and manually compiled a small vocabulary containing less than one thousand words. Based on the vocabulary (i.e., set of terms t_j), a control index provides for each document d_i in the collection (annotations plus document repository), a vector of term scores v_{ij}. We computed the following controlled indices for the document collections (see [17, 11]):

- *Boolean*: the presence of term t_j among the words of document d_i: $v_{ij}^{\text{bool}} = 1$ if $t_j \in d_i$, 0 otherwise
- *Frequency*: the normalized frequency of term t_j in document d_i: $v_{ij}^{\text{freq}} = \frac{f_{ij}}{\max_{\forall j}(f_{ij})}$, in which f_{ij} is the number of occurrences of t_j in d_i
- *tf.idf*: the weighted frequency of term j in document d_i: $v_{ij}^{\text{tf.idf}} = f_{ij} \log(\frac{N}{n_i})$, where N is the total number of documents and n_i is the number of documents containing term i in the collection

Additionally, we computed another type of index called the reference representation (see for example [**?**]). Each annotation contains references to different documents from the repository. As a representation, we consider which documents each annotation refers to:

- *Reference*: The presence of document j as a reference in annotation i: $v_{ij}^{\text{ref}} = 1$ if annotation i contains a ref to document j, 0 otherwise.

3.2 Relevance and similarity metric

To express the similarity between pairs of documents and the similarity of a document to a set of documents, we used the following definitions [17, 11]. For pairs of documents d_i and d_j we used the cosine of the angle between the corresponding normalized vector representations:

$$\text{sim}(d_i, d_j) = \cos(d_i, d_j),$$

denoting the documents and their vector representation by the same symbol. The similarity of a document d_i to a set of documents $C = \{c_1, \ldots, c_L\}$ is defined as

$$g^T(d_i, C) = \frac{1}{L}\sum_{j=1}^{L} \cos(d_i, c_j) + \frac{1}{1 + \text{closeness}(C)},$$

where we use the following definition of closeness for the set of documents C:

$$\text{closeness}(C) = \min_{1 \le j < k \le L} \cos(c_j, c_k).$$

3.3 Pseudo-relevance feedback

Pseudo-relevance feedback methods expand the query with the n most relevant documents in a document collection set [17]. We apply this method by treating the annotations as queries and appending the n most relevant documents from a collection to the annotations (as determined by our document similarity measure). From these expanded annotations, we then regenerate the vector representations described above. In the rest of the paper, we refer to this application of pseudo-relevance feedback as *expansion* (for a related application of the pseudo-relevance feedback with reference representation, see [**?**]). We denote the annotations A expanded with n documents from collection C with A-C_n.

4. CLUSTERING OF YEAST GENES

Although first-generation computational tools for the analysis of expression data are becoming increasingly widespread [16], the assessment of biological meaning to the results constitutes a major challenge. Interpreting cluster patterns involves the consultation of curated functional databases such as Stanford Genome Database[1] (SGD), typically offering a variety of cross-references to other repositories. For even more elaborate information the US National Library of Medicine's MEDLINE provides a common bibliographic source of citations and abstracts in biomedical research from 1966 till present.

The present strategies for knowledge-based expression data analysis rely on the premise that the statistical data analysis and the biological knowledge can complement each other by *linking* two independently constructed sources that contain conceptually related records [12].

Masys *et al.* [5] link groups of genes with relevant MEDLINE abstracts through the PubMed engine[2]. Each cluster is characterized by a pool of the relevant keywords derived from both the MeSH headings and UMLS ontology[3]. The MeSH (Medical Subject Headings) is a controlled vocabulary used for indexing the abstracts in MEDLINE, while the UMLS ontology (Unified Medical Language Systems) is a biomedical concept hierarchy conceived to preserve semantic relations between the concepts described in its controlled vocabularies. Their interface [5] reports the quantitative significance of each result and provides links to different databases to allow further browsing.

Jenssen *et al.* [9] constructed a pioneering online system to link co-expression information from an microarray experiment with their constructed co-citation network. This literature network covers co-occurrence information of gene identifiers in over 10 million MEDLINE abstracts. Their system characterizes co-expressed genes using the MeSH keywords attached to the abstracts about those genes.

Shatkay *et al.* [**?**] link abstracts to genes in a probabilistic scheme that uses the EM algorithm to estimate the parameters of the word distributions underlying a *theme*. Genes are identified as similar when their corresponding gene-by-documents representations are close.

In GEISHA, Blaschke *et al.* [3] profile and evaluate gene clusters by mixing statistical and grammatical analysis (shallow parsing) on PUBMED-retrieved abstracts. GEISHA is based on a comparison of the frequency of abstracts linked to different gene clusters and containing a given term.

We explore the potential and limitations of the vector space model discussed in Section 3, for clustering genes based on their associated literature. To evaluate the biological usefulness of literature clustering, we formulated a clustering problem with gene sets of yeast for which the functional associations are well-established and biologically distinct. The reason not to start immediately from expression-based gene clusters, is that these data-based clusters cannot yet provide a gold standard to interpret and quantify the correspondence between various data mining methods. To compare different versions of the representation with respect to clustering performance, we use an external score for cluster correspondence. The background aim of this evaluation is to establish a powerful statistical text representation as a foundation for the integrated clustering.

4.1 Collection of yeast information

We collected and compiled (Sep 2001) several sources for textual annotations of the genes . Firstly, the Gene Ontology[4] (GO) is a concept hierarchy structured into three main components: molecular function, biological process, and cellular location. Secondly, SWISS-PROT[5] (SP) is a curated protein sequence database. We pooled the GO and SP information into a local database we denote by *YeastCard*. It serves as an extended textual resource for yeast genes. A typical entry is shown in Appendix A.

Finally, as a source for more detailed annotations, we used a collection of 493,923 yeast-related MEDLINE abstracts dated between January 1982 and November 2000. The abstracts originate from 59 journals selected according to their impact factor and their relevance as assessed by a biologist.

We evaluated how these sources can be used for gene clustering and we investigated how the expansion of the GO and YeastCard annotations with MEDLINE abstracts (described in Section 3.3) affect cluster performance.

4.2 Clustering methods

We applied hierarchical clustering and the k-medoids algorithm [10] for the different annotation sources and weighting schemes of the vector space model. k-Medoids takes a variable-to-variable similarity matrix as input and divides the data into k groups by iteratively defining k representative objects (medoids) and reallocating the remaining points to them. As both algorithms use a similarity matrix, we generated such a matrix for each annotation type using the similarity metric outlined in Section 3. We screened the performance of these various annotations by measuring the cor-

[1]http://genome-www.stanford.edu/Saccharomyces/
[2]http://www.ncbi.nlm.nih.gov/PubMed/
[3]http://www.nlm.nih.gov/databases/
[4]http://www.geneontology.org
[5]http://www.expasy.org/sprot/

respondence of the clustering with an external, predefined partition.

As an external score for cluster validity we used the corrected *Rand* index [8]; given a set of n points, an external partition $P = \{P_1, ..., P_k\}$ and a clustering $C = \{C_1, ..., C_l\}$, define a as the number of pairs that occur in the same partition P_i and the same cluster C_j, d as the number of pairs that are grouped differently in P and C, and b and c as the number of pairs that co-occur in P, but not in C or vice-versa. The *Rand* index is then defined by

$$R = \frac{a+d}{a+b+c+d}.$$

The correction for random partitioning is $R_{\text{adj}} = \frac{R-E(R)}{\max(R)-E(R)}$, where a hypergeometric baseline distribution is used to compute the expected values. This yields

$$R_{\text{adj}} = \frac{\sum_i \sum_j \binom{n_{ij}}{2} - \left(\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}\right) / \binom{n}{2}}{\left(\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right)/2 - \left(\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}\right) / \binom{n}{2}}$$

where n_{ij} is the number of elements from P_i that are in C_j, $n_{i\cdot}$ the total number of elements in P_i, and $n_{\cdot j}$ all the elements in C_j. In a comparative study [13], the adjusted *Rand* index is recommended as the external measure of choice.

As an internal score for cluster quality we used the silhouette coefficient $S = \max_k \sum_{i=1}^{n_k} s_{ik}$ where l is the number of found clusters, n_k the size of cluster k and

$$s_{ik} = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where $a(i)$ is the average dissimilarity of member i to all other members of its cluster and $b(i)$ the average dissimilarity of member i to members of nearest cluster [10]. It is a metric-independent measure designed to describe the ratio between cluster coherence and separation and to assist in choosing which clustering is preferable according to the data. We calculated the correlation between the silhouette coefficient and the *Rand* index to evaluate its usefulness in problems where no external assessments are available.

5. BAYESIAN NETWORKS FOR THE PREOPERATIVE ASSESSMENT OF OVARIAN TUMORS

We perform the investigations outlined in Section 2 on a real-world medical problem relating to ovarian cancer. A significant medical goal is to develop mathematical models for the preoperative prediction of the tumor class (e.g., benign vs. malignant). There are two different types of information for the development of such models: the biological and medical information about the disease and the growing amount of patient data. The abundant background knowledge is diverse—for example, the MEDLINE collection of abstracts from biomedical journal papers contains tens of thousands of items about ovarian cancer.

5.1 Domain variables and data

Factors known to affect the risk of malignancy are parity (number of pregnancies), sterility, drug treatment for infertility, duration of lactation, oral contraceptives, foreign body (carcinogens), family history of breast and ovarian cancer, genetic deficiencies, age, age at menopause, hysterectomy, and so on. Additional measurements and observations are the following: bilaterality of the tumor, pelvic pain, morphological descriptors of the mass (such as smoothness and solidness), descriptors of its echogenicity and vascularization, level of several antigens such as CA125, amount of fluid in the abdominal cavity and the day of the cycle. While the effects of some of these variables are well-documented in the literature (such as the effect of the family history and genetic deficiencies), other effects are only qualitatively known and highly subjective (such as the use of the vascularization indices).

In addition to the prior background information, data has been collected in the framework of the IOTA project[6] [19]. The aim of the IOTA project is the prospective collection of data for the development of mathematical models for the preoperative classification of malignant and benign ovarian tumors. The IOTA database contained 68 parameters for 1,150 tumor masses that were used for evaluating the text-based scores for Bayesian network substructures.

5.2 Bayesian networks

A Bayesian network represents a joint probability distribution over a set of variables by exploiting the conditional independence relations. We assume that these variables $V_1, \ldots, V_m$ are discrete and ordered by their index. The model decomposes into a graphical part (a directed acyclic graph) and a numerical part (local dependency models). The vertices of the directed acyclic graph represent the random variables V_i and the edges define the independency relations. Each variable V_i is independent of its non-descendants given its parents, which are denoted as the parental set π_i [15]. There is a local dependency model for each variable to describe its probabilistic dependency on its parents. This decomposed nature of the model induces a two-step procedure for learning. First, the dependency structure is learned (or specified directly by an expert). Second, the parameters for the local dependency models are trained from data. We will focus on the first step and investigate the usage of textual information to perform Bayesian network structure learning in a way similar to the data-driven procedure.

A closed-form Bayesian formula for computing the probability of a Bayesian network structure B_S given a complete data set $\underline{D}$ was derived by Cooper and Herskovits [4]:

$$P(B_S|\underline{D}) \propto P(B_S) \prod_{i=1}^{m} \prod_{j=1}^{q_i} \frac{(r_i-1)!}{(N_{ij}+r_i-1)!} \prod_{k=1}^{r_i} N_{ijk}!.$$

In the formula, the first product goes over all domain variables. The second product iterates over all q_i *different* configurations for the parents π_i of variable V_i that are found in the data set. The last product iterates over the r_i possible values for variable V_i. The quantities N_{ijk} contain the number of times we observe a value k for the i-th variable while its parents are at the j-th parental configuration and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ (for details, see [4]).

5.3 A data-based local dependency score

Note that the probability of a Bayesian network structure given a complete data set can be decomposed into a product of independent parts, which we define to be $g^D(V_i, \pi_i)$, each expressing the probability of the local dependency model of

[6]https://www.iota-group.org

variable V_i with parents π_i conditioned on the data:

$$P(B_S|\underline{D}) = P(B_S)\prod_{j=1}^{n} g^D(V_i, \pi_i).$$

Despite the decomposition of the learning to the selection of appropriate parental sets, the amount of data needed for statistically significant identification of networks is still considerably high. One potential solution is to define an informative *a priori* distribution over all possible local dependency structures and update this distribution to an *a posteriori* distribution using the data. Because of the exponential number of possible sets, this task is very difficult for human experts (for certain methods, see [7]). A natural step would be to use textual information for the definition of this *a priori* distribution over the local dependency structures. To investigate the feasibility of such conversion we compare the previous data-based score $g^D(V_i, \pi_i)$ for these local substructures with a newly defined text-based score $g^T(V_i, \pi_i)$.

5.4 Annotated Bayesian networks

A recent extension of the Bayesian network representation of probability distributions, the Annotated Bayesian Network (ABN), allows the attachment of free text to the objects in the representation [1]. On the one hand, the ABN can be useful to document the incorporated heterogeneous sources of information in the network. On the other hand, it provides a formal framework in which the probabilistic, computational model is linked with textual resources (i.e., it provides a framework to investigate the potential of how the textual knowledge can be used in the model building and identification). In the manual case, the ABN serves as the context of the modeler for information retrieval while building the model [2]. In the automatic case, which we are investigating in this paper, the incorporated textual information supports structure learning based on the scores introduced below. This annotated Bayesian network representation and a corresponding implemented system provided the formal framework and the experimental environment for investigation of the text and data-based scores for Bayesian network substructures.

5.5 A text-based local dependency score

For simplicity, we denote the domain variable, the stochastic variable, the annotation of a stochastic variable, and the corresponding vector representation of the text identically. Using the definitions from Section 3.2 (3.2), the text-based score for variable V_i and parental set π_i is defined by $g^T(V_i, \pi_i)$ (with this notation, we therefore mean the document similarity between the annotation of variable V_i and the set of annotations of its parents π_i). This score characterizes the mean distance between a child variable V_i and the parent variables in the set π_i, additionally it penalizes if the parent annotations are too similar.

5.6 Source of annotations

The twenty-five page IOTA protocol used for data collection is the primary source of the annotations and contains: (1) general information about the project, (2) inclusion and exclusion criteria for patient records, (3) a description of each variable with its format, its value list, mandatory/optional constraints and possible inter-variable dependency rules, (4) the grouping of variables into sections and (5) the diagnostic methods for all tumor variables together with self-explaining figures. A corresponding Ph.D. thesis [18] and The Merck Manual[7] provided an extension for the IOTA descriptions. Together, these compose one type of annotations, the *free-text* annotation (T), on average a hundred-word description for each of the twenty-six domain variables.

Another type of annotation, the *manual references* (R), was derived by asking two experts to select electronically available medical references for the variables that are most relevant in the IOTA context. They selected forty-two and twenty-two separate references, which are attached in a non-exclusive way to the domain variables, on average three to five references for the eighteen variables that are covered.

Additionally we asked the experts to select journals as most relevant for the domain (2 journals), highly relevant (3 journals), moderately relevant (33 journals), and relevant journals (93 journals). We constructed four collections of MEDLINE abstracts containing 5,367, 71,845, 231,582 and 378,082 abstracts selected from the MEDLINE corpus dated between January 1982 and November 2000. These collections were used to select the most relevant MEDLINE items and automatically expand the annotations (denoted by ML_j^i if the ith collection is used to select j number of items).

Finally, we constructed another collection to investigate the effect of expansion. This is based on the On-line Medical Dictionary[8] and the CancerNet Dictionary[9]. In total it contains 67,829 short entries. The expansion with j items from this collection is denoted by O_j.

A typical entry composed of these sources is shown in the Appendix B.

6. RESULTS

We now present for the different textual information sources on the two problems of clustering and substructure learning for Bayesian networks.

6.1 Clustering

We constructed a set of three groups for which the functional associations are well-established. The first group contains 63 genes that encode lysosomal proteins, the second contains 30 genes involved in translational control, and the third contains 23 genes related to amino acid transport. For all these genes we selected their corresponding GO and YeastCard annotations (see Section 4.1) and represented them by the vector schemes outlined in Section 3. Furthermore, we expanded these annotations with the 20 best matching MEDLINE abstracts, indicated by GO-ML_{20} and YC-ML_{20} respectively. These expansions were indexed according to various indexing schemes. Next, we clustered the different textual gene profiles setting the number of clusters to three. Table 1 lists R_{adj} for the most important combinations of annotation, representation and clustering method. Among the hierarchical clustering methods (single, complete, and average linkage and Ward's method) , Ward's method proved the only reasonable one. Furthermore, k-medoids generally outperformed the hierarchical method as can be observed in columns *Hier* and *KMed* of Table 1. In the remainder of this section we will therefore only refer to

[7]http://www.merck.com/pubs/mmanual
[8]http://www.graylab.ac.uk/omd/index.html
[9]http://thymoma.de/meddict.htm

values in the *KMed* column.

In our analysis, GO (which provides only brief keyword annotations) does not provide sufficient information for an acceptable statistical representation. Our compiled YeastCard annotation indicated as *YC tf.idf*, has R_{adj} of 0.4698, which is much better than the score of 0.1608 for GO alone. Expanding the GO modifies R_{adj} from 0.1608 to 0.5792. An expansion of YeastCard on the other hand further improves the score from 0.4698 to 0.6948. We observe that the score of expanding the textually richer YeastCard ($R_{\text{adj}} = 0.6948$) is, in turn, higher than that of the GO-based expansion ($R_{\text{adj}} = 0.5792$). It shows that richer annotations yield better expansions.

Table 1 also demonstrates how different representations affect cluster effectiveness. For the GO, the boolean representation is most suited among the options (results not shown). The use of a stopword list, indicated as *GO bool restr* in Table 1, attempts to eliminate possibly distorting words as *unknown* and *null*, but shows no improvement in the GO case. When textual descriptions become larger than approximately 100 words, as is the case with the YeastCard and the expansions, we found the boolean representation to perform worse than the frequency-based representations *freq* ($R_{\text{adj}} = 0.6032$) and *tf.idf* ($R_{\text{adj}} = 0.5792$). When dealing with the MEDLINE expansions, the reference representation (*ref repr*) scores significantly less ($R_{\text{adj}} = 0.2354$) than the alternatives.

In Table 2 we print the contingency table for the best annotation and representation. It shows the correspondence between the clustering and the external grouping for $R_{\text{adj}} = 0.6948$.

Finally, the correlation between the R_{adj} and S is 0.0457. To gain more insight into the discrepancy between the internal and external index, we examined how the silhouette scores per cluster, $s^r_{\cdot k} = \frac{1}{n_r}\sum_{i=1}^{n_r} s_{ik}$, change in function of the n_r genes that lie closest to their medoid. In Fig. 1 we plot such a silhouette profile for the clusters computed for the YeastCard expansion. The flat regions indicate that no genes are present in the respective dissimilarity range, while sudden drops in silhouette scores show the detrimental effect of a set of distant genes on the silhouette score. The overall quality of the text representation (which is determined by the quality of the text source, the preprocessing steps, the retrieval process, and the ability of the vector representation to encode real-world concepts) will influence the correlation between the external and internal scores directly.

Table 1: Adjusted *Rand* scores for different annotations, representations, and clustering methods.

Annotation	Weighting scheme	*Hier*	*KMed*
GO	*bool*	0.2494	**0.1608**
GO	*bool restr*	0.2252	0.1561
GO-ML$_{20}$	*bool*	0.3391	0.4177
GO-ML$_{20}$	*freq*	0.4476	0.6032
GO-ML$_{20}$	*tf.idf*	0.2997	**0.5792**
GO-ML$_{20}$	*reference*	0.2364	0.2354
YC	*bool*	0.2159	0.0805
YC	*freq*	0.2752	0.2710
YC	*tf.idf*	0.3446	**0.4698**
YC-ML$_{20}$	*tf.idf*	0.3988	**0.6948**

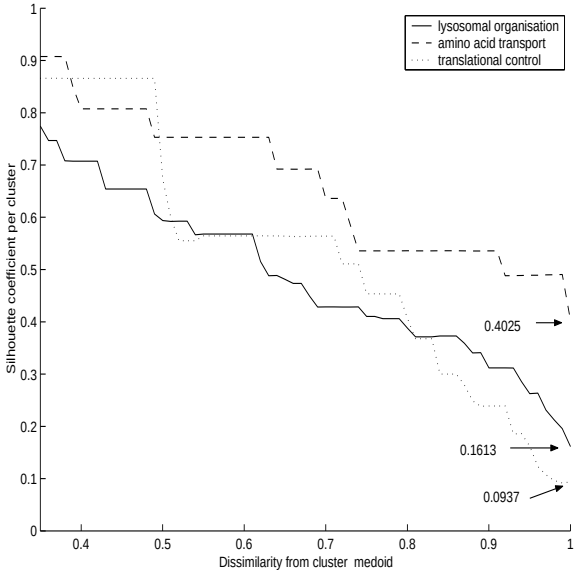

Figure 1: Effect of distant members in each cluster on its silhouette score: starting with the nearest members ($1 - \text{sim}(\text{member}, \text{medoid}) < 0.35$), we gradually monitor changes in the silhouette score per cluster, $s^r_{\cdot k}$, by including increasingly distant members. Flat regions indicate that no genes are present in the respective dissimilarity range. Sudden drops in silhouette scores show the detrimental effect of those more distant genes on the scores. The overall silhouette coefficient $S = 0.2192$ is the mean of silhouette scores per cluster, which are indicated by the arrows.

Table 2: Contingency table for best clustering.

	C_1	C_2	C_3
P_1	45	7	0
P_2	2	28	0
P_3	0	2	20

6.2 Evaluation of text-based score for Bayesian networks

To investigate the possibility of using integrated text- and data-based scores for learning Bayesian networks, we compared the text-based scores introduced in Section 5.5 against (1) the prior domain knowledge and (2) the data-based scores. The available prior knowledge in the ovarian cancer domain consists of a ranking by the expert of the domain variables according to their relevance for discriminating between benign and malignant tumors. This ranking thus represents here an assessment of the relevance of each domain variable for predicting the *Pathology* variable. On the one hand, we can compare the expert's ranking to a *data*-based ranking of the domain variables. This is obtained by the data-based scores $g^D(\text{Pathology}, \pi_{\text{Pathology}})$ for pairs of the *Pathology* variable and the remaining domain variables (i.e., parental sets of size 1). On the other hand, the expert's ranking can be compared against a *text*-based ranking of the domain variables based on the text-based scores $g^T(\text{Pathology}, \pi_{\text{Pathology}})$. Table 3 presents the rankings of

the domain variables by their relevance to the *Pathology* variable according to a medical expert, the statistical data, and the literature.

Table 3: Relevance ranking of variables for the variable *Pathology* by text, data, and a medical expert

Rank	Text	Data	Expert
1	ColScore	CA125	CA125
2	CA125	Ascites	ColScore
3	Locularity	ColScore	Papillation
4	Volume	WallRegularity	Volume
5	Papillation	Locularity	Ascites
6	Septum	Volume	Age
7	PMB	Age	Bilateral
8	Pregnant	Shadows	Locularity
9	Echogenicity	Papillation	Shadows
10	WallRegularity	Bilateral	-
11	Origin	Septum	-
12	Age	Meno	-

In Fig. 2 the domain variables are positioned on the coordinates $(g^T(\text{Pathology}, V_i), g^D(\text{Pathology}, V_i))$ to illustrate further the correlation between text- and data-based scores. Fig. 3 shows all pairwise relevance scores $g^T(V_i, V_j)$.

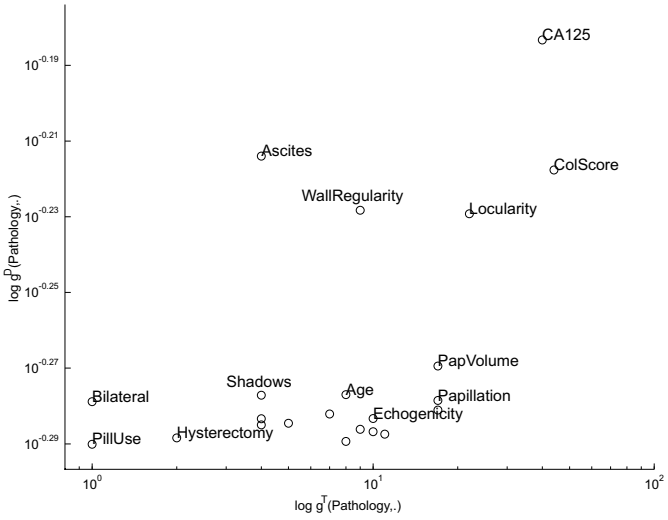

Figure 2: Text- and data-based relevance scores for the domain variables and *Pathology* (*free text* annotation, Boolean representation, the small domain vocabulary).

To evaluate the effect of the different text representations, annotations type, and domain vocabularies on the relation of text and data scores, we computed the correlation coefficient and the Spearman rank correlation coefficient R_S. For the variable V_i, R_S is defined as

$$R_S = 1 - 6\frac{\sum_{j=1}^{P_i}(\text{Rank}_{\text{Text}}(\pi_{ij}) - \text{Rank}_{\text{Data}}(\pi_{ij}))^2}{P_i(P_i^2 - 1)},$$

where P_i is the number of possible parental sets for variable V_i and $\pi_{ij}, j = 1, \ldots, P_i$ are all the possible parental sets (which are all possible combinations of the other variables upto a certain fixed number of parents t, i.e., we have $P_i = \binom{t}{n-1}$ possible combinations).

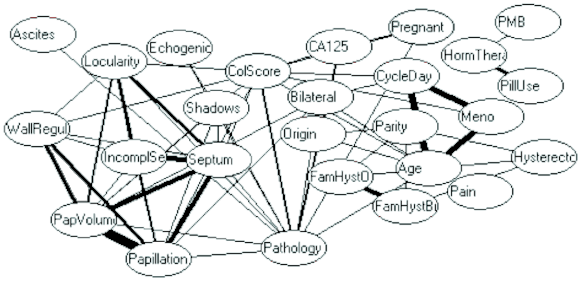

Figure 3: The vizualisation of text-based relevance scores $g^T(V_i, V_j)$ for the pairs of domain variables (*free text* annotation, *tf.idf* representation, the small domain vocabulary, threshold 0.2).

We also report the average of the Spearman rank correlation coefficients for the variables. Additionally, we report a special rank-correlation measure defined as follows. For each variable, the text- and data-based scores define a text rank and data rank for the parental sets. Define a matrix R in which the a_{kl} element is the number of times the parental sets have text rank k and data rank l. Clearly, if the scores or their rankings for each variable are identical this will be a diagonal matrix. Now define a matrix R' which is the 4-by-4 partitioning of R with the following intuitive interpretation for the four partitions: highly relevant, moderately relevant, less relevant, and not relevant. The respective diagonal consists the following pairs from upper left to lower right: (highly relevant by text, highly relevant by data), (moderately relevant by text, moderately relevant by data), and so on. We report the normalized trace of R', that is the correspondence between the text and data-based ranking using this 4-graded granularity for all the variables and only for *Pathology* also.

Table 4 presents the results for the most interesting settings while Table 5 contains a more structured and detailed reports for a larger number of settings.

Table 4: Relations between text- and data-based scores. The correlation coefficients, the Spearman rank correlations, and the normalized trace of the text-rank–data-rank matrices are reported for the respective settings (*free text*, optionally expanded with dictionary entries or MEDLINE abstracts, set size is 1).

	For *Pathology*			For all variables	
Settings	Corr	Trace	R_S	$\widehat{\text{Trace}}$	$\widehat{R_S}$
bool, T	0.73	0.52	0.69	0.40	0.34
tf.idf, T	0.69	0.44	0.81	0.36	0.33
tf.idf, T-O_3	0.49	0.44	0.80	0.34	0.31
bool, T-ML^4_{12}	0.71	0.48	0.61	0.34	0.30

7. DISCUSSION

In our study on text clustering, we found the *freq* and *tf.idf* applied on the expanded annotation to be superior to the

boolean representation and the reference representation. An optimal choice between them, however, depends on the annotation source and cannot be known in advance. The Gene Ontology and functional annotation database SWISS-PROT proved valuable sources of free-text information, especially if used as a query in the expansion step. This illustrates that curated databases of structured and unstructured information not only provide indispensable access to information, but also constitute useful sources to automatically extract knowledge from domain literature. The poor performance of the reference representation partly can be explained by its sensitivity to the quality of the expansion (kernel quality and retrieval quality).

Because no data set of gene expression data can serve as a high-quality benchmark for clustering, we conducted our comparison of various annotations and vector representations on a custom gene partition. Although the constructed clustering problem is fairly easy from a biological viewpoint, it made it possible to isolate the effects of various information sources and parameterizations on the cluster performance. We found that internal scores can provide an important confidence measure in the quality of text-clustering and indirectly in the comparison between text-based and data-based clustering.

One of our aims is to construct a statistical representation suitable for integrating prior knowledge in expression-based gene clustering, We therefore outline in Fig. 7 how we plan to use the current representation. Using the terminology in [14], we depict early, intermediate, and late integration of expression data and text. Early integration pools both types of statistical data and passes it to the cluster algorithm. Intermediate integration creates one variable-to-variable similarity matrix for each data type, merges them in some way, and passes them to a clustering algorithm. Finally, late integration compares or merges two separate analyses. The question which of these schemes provide a good foundation for integrated cluster analysis constitutes a topic of our future research.

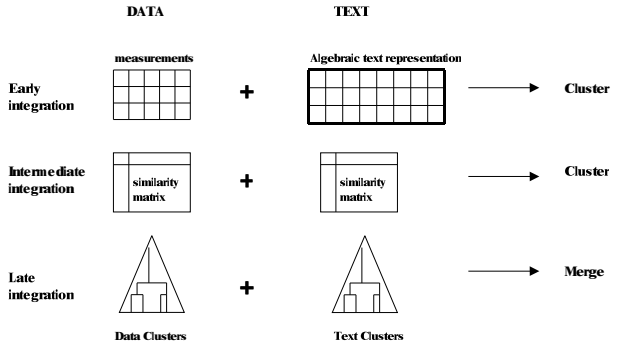

Figure 4: Various ways of integrating domain literature and data in clustering

For the Bayesian network, the text- and data-based scores proved to be significantly correlated and rank-correlated. From the medical point of view, the ranking of the parental sets seems surprisingly good. Contrary to our expectations, the expansion of the annotations with manual or automatically selected references could not improve the performance, which needs further investigations. The related reference representation similarly has a poor performance.

One future research direction is the incorporation of text-based scores in various Bayesian network algorithms. The Bayesian framework presented in Section 5.2 offers a principled method for the incorporation of prior domain knowledge through prior distributions. Currently, we are investigating methods for such a transformation and evaluate their effects on the classification performance in the ovarian cancer domain. A related direction is the general investigation of the score $g^T(.,.)$ and the corresponding conditional independence statements in an induced Bayesian network.

Finally, in both applications the text-based scores, that is sim(.,.), currently relies on a vector representation of the text, while the annotations are already structured into various fields. For example, the MEDLINE references contain manually curated MeSH keywords. A structured text representation with a corresponding text-based score exploiting this structural information may have many advantages. A related research topic would be a more refined linguistic analysis, including better phrase identification or shallow parsing for a better representation of the content yielding an improved sim(.,.), enhancing also the text-based score.

8. CONCLUSION

In this paper, we assumed to have short free-text descriptions for the domain variables and a huge repository of related domain literature. We used data and external assessments for evaluation purposes. We considered the problem of identifying which text representations and statistical scores best support the use of literature in statistical models. We investigated this potential for two statistical methods: clustering and Bayesian network learning. Firstly, we reported the performance in clustering yeast genes against an expert reference. Secondly, we reported the correspondence between text and data in scoring Bayesian network substructures in the medical task of modeling the joint distribution of clinical measurements of ovarian tumors. Results reported for various types of textual information sources and vectorial text representations indicate that the use of literature and statistical data can be formulated in a common framework and their effects can be compared. This suggests that closely coupled representations and methods are a viable foundation in the development of integrated text and data analysis methods.

9. ACKNOWLEDGMENTS

Dr. Bart De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium. Yves Moreau is an assistant professor at the K.U.Leuven and a postdoctoral researcher with the Belgian Fund for Scientific Research (F.W.O. - Vlaanderen). Peter Antal and Patrick Glenisson are Research Assistants with the K.U.Leuven. Geert Fannes is a Research Assistant with the F.W.O. Vlaanderen. Our research is supported by grants from: Research Council KUL: Concerted Research Action GOA-Mefisto 666, IDO (IOTA Oncology, Genetic networks); Flemish Government: Fund for Scientific Research Flanders (projects G.0256.97, G.0115.01, G.0240.99, G.0197.02, G.0407.02, research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/ Poland), IWT (Soft4s, STWW-Genprom, GBOU-McKnow, Eureka-Impact, Eureka-FLiTE); Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006): Dynamical Systems and Control: Computation, Identification &

10. ADDITIONAL AUTHORS

Janick Mathys (KUL/ESAT-SCD,
email: janick.mathys@esat.kuleuven.ac.be)
and Bart De Moor (K.U.Leuven/ESAT-SCD,
email: bart.demoor@esat.kuleuven.ac.be)
and Yves Moreau (K.U.Leuven/ESAT-SCD,
email: yves.moreau@esat.kuleuven.ac.be).

11. REFERENCES

[1] P. Antal, T. Meszaros, B. D. Moor, and T. Dobrowiecki. Annotated bayesian networks: a tool to integrate textual and probabilistic medical knowledge. In *Proc. of the 13th IEEE Symp. on Comp.-Based Med.Sys. (CBMS01)*, pages 177–182, 2001. Bethesda, MD.

[2] P. Antal, T. Meszaros, B. D. Moor, and T. Dobrowiecki. Domain knowledge based information retrieval language: an application of annotated bayesian networks in ovarian cancer domain. In *Proc. of the 14th IEEE Symp. on Comp.-Based Med.Sys. (CBMS02)*, pages .–., 2002. Maribor, Slovenia.

[3] C. Blaschke, J. Oliveros, and A. Valencia. Mining functional information associated with expression arrays. *Funct Integr Genomics*, 1:256–268, 2001.

[4] G. F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[5] D. M. et al. Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, 17:319–326, 2001.

[6] W. B. Frakes. *Stemming algorithms*. in W. B. Frakes and R. Baeze-Yates: Information retrieval. Prentice Hall, 1992.

[7] D. Heckerman. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[8] A. Jain and R. Dubes. *Algorithms for clustering data*. Prentice Hall, 1988.

[9] T. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28:21–28, may 2001.

[10] L. Kaufman and P. Rousseeuw. *Finding groups in data*. Wiley-Interscience, 1990.

[11] R. Korfhage. *Information Storage and Retrieval*. New York: Wiley Computer Pub., 1997.

[12] D. Masys. Linking microarray data to the literature. *Nature Genetics*, 28:9–10, 2001.

[13] G. Milligan and M. Cooper. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavorial Research*, 21:441–458, 1986.

[14] P. Pavlidis, J. Weston, J. Cai, and W. Grundy. Gene functional classification from heterogeneous data. In *Proceedings of the 5th International Conference on Computational Molecular Biology (RECOMB 2001)*, 2001.

[15] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1989.

[16] J. Quakenbush. Computational analyis of microarray data. *Nature Reviews Genetics*, 2:418–427, 2001.

[17] B. R.-N. R. Baeza-Yates. *Modern Information Retrieval*. ACM Press, 1999.

[18] D. Timmerman. *Ultrasonography in the assessment of ovarian and tamoxifen-associated endometrial pathology*. Ph.D. dissertation, Leuven University Press, D/1997/1869/70, 1997.

[19] D. Timmerman, L. Valentin, T. H. Bourne, W. P. Collins, H. Verrelst, and I. Vergote. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the international ovarian tumor analysis (iota) group. *Ultrasound Obstet Gynecol*, 16(5):500–505, Oct 2000.

APPENDIX

A. AN ENTRY IN YEASTCARD

A typical entry in our local database *YeastCard* is composed of Gene Ontology and SWISS-PROT information.

- *Gene Name*: YER133W
- *GO Biological Process*: protein phosphatase type1
- *GO Molecular Function:* glycogen metabolism
- *GO Description*: Glycogen accumulation
- *SP Protein Name*: cytoplasmic
- *SP Function*: involved in control of glycogen metabolism meiosis translation chromosome segregation cell polarity and g2/m cell cycle progression. pp1 may act in opposition to the ipl1 protein kinase in regulating chromosome segregation.
- *SP SubCellular Location*: cytoplasmic
- *SP Similarity*: belongs to the ppp family of phosphatases. pp-1 subfamily.
- *SP Organism*: yeast, saccharomyces cerevisiae

B. ANNOTATION FOR A BAYESIAN NETWORK VARIABLE

A typical annotation for a domain variable is illustrated with the annotation of the variable CA125: "The ca 125 antigen is a glycoprotein that is expressed by most epithelial ovarian and is recognized by a monoclonal antibody. serum ca 125 is the tumor marker with the highest sensitivity for ovarian cancer (bast et al 1983, jacobs et al 1989, knapp et al 1996, cuckle 1996). This tumor marker will detect nearly 80 percent of advanced (stage iii) ovarian cancers (see table i), but only about 44 percent of patients with stage i disease (vergote et al 1987, cuckle & wald 1991, bourne et al 1994a, maggino et al 1994). In premenopausal patients the specificity is low at a cut-off level of 35 u/ml, since false positive results are frequently encountered in menstruating or pregnant women and in a wide variety of benign conditions, such as benign ovarian tumors (10%) endometriosis (20-30%), liver cirrhosis (60-70%), pancreatitis (30%), pelvic inflammatory disease, uterine fibroids and meigs syndrome (100%).· · · "

Table 5: Relation between text and data scores. The different columns contain (1) the vector representation, (2) the source of annotation, (3) the number of parents, (4) the vocabulary, (5) the correlation coefficient between the text score and the data score for all parental configurations of the *Pathology* variable, (6) the normalized trace of the 4-by-4 partition of the text-rank–data-rank matrix, (7) the Spearman correlation coefficient for the parental text–data scores of the *Pathology* variable, (8) the Z-score of the hypothesis test that there is no monotonic relationship between the text and data ranks, (9) the probability of the hypothesis test that there is no monotonic relationship between the text and data ranks, (10) the normalized trace of the 4-by-4 partition of the text-rank–data-rank matrix for *all* variables, and (11) the average Spearman correlation coefficient for the parental text–data scores of the *all* variables.

Repr	Annot	#II	Voc	$\text{Corr}^{\text{Path}}$	$\text{Trace}^{\text{Path}}$	R_S^{Path}	Z	Probability	$\text{Trace}^{\text{all}}$	R_S^{all}
tf.idf	T	1	restr.	0.691975	0.44	0.810769	3.97194	7.12897e-005	0.367692	0.338432
tf.idf	T	1	all	0.691975	0.44	0.810769	3.97194	7.12897e-005	0.364615	0.33855
bool	T	1	restr.	0.735572	0.52	0.695385	3.40667	0.000657606	0.404615	0.346124
bool	T	1	all	0.735572	0.52	0.695385	3.40667	0.000657606	0.404615	0.34574
freq	T	1	restr.	0.369573	0.4	0.730769	3.58002	0.000343568	0.276923	0.181538
freq	T	1	all	0.369573	0.4	0.730769	3.58002	0.000343568	0.28	0.182544
tf.idf	T	2	restr.	0.54503	0.4	0.648474	11.2132	3.5126e-029	0.34359	0.295509
tf.idf	T	2	all	0.54503	0.4	0.648474	11.2132	3.5126e-029	0.34359	0.295509
bool	T	2	restr.	0.714902	0.503333	0.65955	11.4047	3.96141e-030	0.361538	0.296987
bool	T	2	all	0.714902	0.503333	0.65955	11.4047	3.96141e-030	0.361538	0.296987
freq	T	2	restr.	0.281367	0.376667	0.437684	7.56826	3.78256e-014	0.283718	0.131722
freq	T	2	all	0.281367	0.376667	0.437684	7.56826	3.78256e-014	0.283718	0.131722
tf.idf	TR	1	restr.	0.462424	0.4	0.621538	3.0449	0.00232758	0.316923	0.276953
bool	TR	1	restr.	0.00520293	0.28	-0.140769	-0.689626	0.490429	0.301538	0.178491
tf.idf	TR	2	restr.	0.296096	0.313333	0.406099	7.0221	2.18558e-012	0.350128	0.272022
bool	TR	2	restr.	0.00342128	0.203333	-0.114168	-1.97415	0.0483647	0.336154	0.240664
tf.idf	T-O_3	1	restr.	0.499581	0.44	0.808462	3.96064	7.47492e-005	0.344615	0.311479
bool	T-O_3	1	restr.	0.711297	0.48	0.613077	3.00345	0.00266937	0.356923	0.30855
tf.idf	T-O_3	2	restr.	0.395009	0.41	0.55675	9.62711	6.14329e-022	0.45359	0.383198
bool	T-O_3	2	restr.	0.694253	0.48	0.629456	10.8843	1.36947e-027	0.469231	0.424064
tf.idf	T-ML_{12}^0	1	restr.	0.499581	0.44	0.803077	3.93426	8.34534e-005	0.343077	0.287811
bool	T-ML_{12}^0	1	restr.	0.711297	0.48	0.613077	3.00345	0.00266937	0.343077	0.304586
tf.idf	T-ML_{12}^0	2	restr.	0.395009	0.41	0.55675	9.62711	6.14329e-022	0.45359	0.383198
bool	T-ML_{12}^0	2	restr.	0.694253	0.48	0.629456	10.8843	1.36947e-027	0.469231	0.424064
tf.idf	T-ML_3^3	1	restr.	0.499581	0.44	0.806923	3.9531	7.71452e-005	0.330769	0.300976
bool	T-ML_3^3	1	restr.	0.711297	0.48	0.616923	3.02229	0.0025087	0.321538	0.279734
tf.idf	T-ML_3^3	2	restr.	0.395009	0.41	0.55675	9.62711	6.14329e-022	0.45359	0.383198
bool	T-ML_3^3	2	restr.	0.694253	0.48	0.629456	10.8843	1.36947e-027	0.469231	0.424064