Dynamic Hybrid Clustering of Bioinformatics by Incorporating Text Mining and Citation Analysis

Frizo Janssens
Electrical Engineering (ESAT)
Katholieke Universiteit Leuven
Kasteelpark Arenberg 10
B-3001 Leuven (Belgium)
frizo.janssens@esat.kuleuven.be

Wolfgang Glänzel
Steunpunt O&O Indicatoren
Katholieke Universiteit Leuven
Dekenstraat 2
B-3000 Leuven (Belgium)
wolfgang.glanzel@econ.kuleuven.be

Bart De Moor
Electrical Engineering (ESAT)
Katholieke Universiteit Leuven
Kasteelpark Arenberg 10
B-3001 Leuven (Belgium)
bart.demoor@esat.kuleuven.be

ABSTRACT

To unravel the concept structure and dynamics of the bioinformatics field, we analyze a set of 7401 publications from the Web of Science and MEDLINE databases, publication years 1981–2004. For delineating this complex, interdisciplinary field, a novel bibliometric retrieval strategy is used. Given that the performance of unsupervised clustering and classification of scientific publications is significantly improved by deeply merging textual contents with the structure of the citation graph, we proceed with a hybrid clustering method based on Fisher's inverse chi-square. The optimal number of clusters is determined by a compound semi-automatic strategy comprising a combination of distance-based and stability-based methods. We also investigate the relationship between number of Latent Semantic Indexing factors, number of clusters, and clustering performance. The HITS and PageRank algorithms are used to determine representative publications in each cluster. Next, we develop a methodology for dynamic hybrid clustering of evolving bibliographic data sets. The same clustering methodology is applied to consecutive periods defined by time windows on the set, and in a subsequent phase chains are formed by matching and tracking clusters through time. Term networks for the eleven resulting cluster chains present the cognitive structure of the field. Finally, we provide a view on how much attention the bioinformatics community has devoted to the different subfields through time.

Categories and Subject Descriptors

I.5.3 [**Pattern Recognition**]: Clustering; H.2.8 [**Database Management**]: Database Applications—*scientific databases, data mining*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms

KDD'07, August 12–15, 2007, San Jose, California, USA.

Keywords

Fisher's inverse chi-square method, cluster chains

1. INTRODUCTION

Bioinformatics is an interdisciplinary field that emerged from the increasing use of computer science and information technology for solving problems in biomedicine, mostly at the molecular level. *Ouzounis* and *Valencia* have provided a review of the early stages of the long history of bioinformatics [19]. In recent studies by *Patra* and *Mishra* [20] and *Perez-Iratxeta et al.* [21], evolution and trends in bioinformatics research have been studied. The field has been characterized as an emerging, dynamically evolving discipline with astonishing growth dynamics. The studies were based on the MEDLINE database and partially on NIH-funded project grants. In both cases, bioinformatics was analyzed in a broad biomedical context.

In a recent study, the authors have analyzed the bioinformatics field from a bibliometric point of view [7], including patterns of national publication activity, citation impact, and international collaboration. A novel subject-delineation strategy was developed for the retrieval of the core literature in bioinformatics.

The goal of the present paper is to investigate the cognitive structure and the dynamics of this *core* of bioinformatics and of its sub-disciplines, based on information from the *ISI* Web of Science database and MEDLINE. We apply techniques from text mining, Web mining and bibliometrics. Text mining and citation analysis both provide effective and valuable algorithms for mapping of knowledge and for charting science and technology fields. The textual and graph-based approaches provide different perceptions of similarity between documents or groups of documents. We incorporate both viewpoints since an integrated approach leads to a better comprehension of the structure of bibliographic corpora [13] (p. 130–152). The actual integration is achieved by statistical combination of various distances, between the same pair of documents, but stemming from different dissimilarity measures exploiting different views on the documents. The integration method based on Fisher's inverse chi-square has previously been shown to significantly outperform corresponding text-only and link-only methods, as well as other integration schemes [13]. For the present analysis, we use this hybrid hierarchical clustering algorithm to combine bibliographic coupling [14] with text-based similarities.

Section 2 briefly describes the delineation of the bioinformatics data set by use of the *bibliometric retrieval* strategy. Next, the Methods section discusses the representation of textual data (3.1), citation analysis (3.2), and clustering (3.3), including the determination of the optimal number of clusters. Procedures for hybrid clustering by weighted linear combination of distance matrices and by Fisher's inverse chi-square method are explained in Section 3.4. Then, in Section 3.5, we introduce *dynamic hybrid clustering* by matching and tracking clusters through time. Results for the hybrid and for the dynamic hybrid analyses are given in Sections 4 and 5. In Section 4.1.1, we also investigate the relationship between number of Latent Semantic Indexing factors, number of clusters, and clustering performance. Finally, the *cluster chains* and their structure and evolution are analyzed in Sections 5.1, 5.2, and 5.3.

2. DATA SET

In a forthcoming study, a novel subject delineation strategy has been developed for retrieving the *core* literature in bioinformatics [7]. It is a combination of textual components and bibliometric, citation-based techniques. The data set resulting from this *bibliometric retrieval* strategy was extracted from the Web of Science Edition of the *Science Citation Index Expanded*^{TM} of *Thomson Scientific* (Philadelphia, PA, USA) and consists of 7401 bioinformatics-related *articles*, *notes*, and *reviews*, published between 1981 and 2004. Each included paper has also been matched against MEDLINE in order to retrieve associated Medical Subject Headings (MeSH). From each record we considered the textual information present in titles and abstracts, author keywords, and MeSH terms. In addition, we collected all cited references and all citing papers.

Figure 1 shows the yearly number of publications in the bioinformatics set. Seven periods are defined for the dynamic analysis in Section 3.5 and Section 5, while striving for comparable numbers of publications in each period.

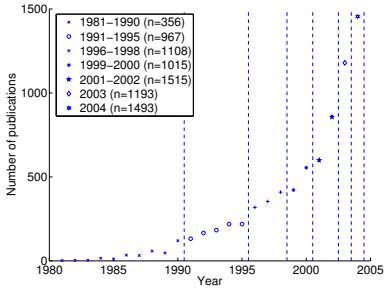


Figure 1: Evolution of publication output in bioinformatics. Time windows for the dynamic analysis are indicated with vertical lines as well as in the legend.

The bibliometric study by *Patra* and *Mishra* was based on MeSH terms and adopted a rather liberal delineation strategy that was tailored towards maximal recall. They selected 14.563 journal articles, that is, about twice as many as we have found [20]. The main reason is the broad interpretation of bioinformatics resulting from the less restricted search strategy. The other reason is the broader coverage of the underlying database.

3. METHODS

3.1 Text analysis

All textual content was indexed with the Jakarta Lucene platform [9] and encoded in the Vector Space Model using the TF-IDF weighting scheme [1]. Text-based similarities were calculated as the cosine of the angle between the vector representations of two papers [23]. *Stop words* were neglected during indexing and the *Porter* stemmer was applied to all remaining terms from titles and abstracts. Bigrams were detected from a candidate list of MeSH descriptors, author keywords, and noun phrases. The dimensionality of the term-by-document matrix was reduced from 18.163 term dimensions to 10 factors by Latent Semantic Indexing (LSI) [6, 4]. LSI is a dimensionality reduction technique based on the Singular Value Decomposition (SVD) of a term-by-document matrix. In Section 4.1.1, we demonstrate that this LSI with only 10 factors provides a local maximum in clustering performance.

3.2 Citation analysis

Important and highly recognized bioinformatics papers can be identified by analyzing the citation graph. We use the link-based algorithms HITS [15] and PageRank [5] to determine representative publications (see Table 1).

The cosine measure used to quantify the text-based similarity between any two documents can analogously be used with Boolean input vectors indicating the cited references in an article, or indicating all citing articles. This corresponds to bibliographic coupling (BC) [14] and co-citation, respectively, which are two citation-based measures of similarity. In the present study we use bibliographic coupling and combine it with text-based similarities in order to obtain an integrated measure that can be used by a clustering algorithm.

3.3 Clustering

To subdivide the bioinformatics papers into clusters we used the agglomerative hierarchical clustering algorithm with Ward's method [12]. It is a hard clustering algorithm, which means that every publication is assigned to exactly 1 cluster.

Optimal number of clusters

Determination of the optimal number of clusters in a data set is a difficult issue and depends on the adopted validation and chosen similarity measures, as well as on data representation. The strategy that we used to determine the optimal number is a combination of distance-based and stability-based methods. This compound strategy encompasses observation of a dendrogram, text- and citation-based mean Silhouette curves, and a stability diagram.

Dendrogram.

A preliminary judgment is offered by a dendrogram, which provides a visualization of the distances between (sub-)clusters. It shows the iterative grouping or splitting of clusters in a hierarchical tree (see Figure 6 for an example). A candidate optimal number of clusters can be determined visually by looking for a cut-off point where an imaginary vertical line would cut the tree such that resulting clusters are well separated. Because of the difficulty to define the optimal cut-off point on a dendrogram [12], we complement this method with other techniques.

Silhouette curves.

A second appraise for the optimal number of clusters is given by the mean Silhouette curve. The Silhouette value for a document ranges from -1 to +1 and measures how similar it is to documents in its own cluster vs. documents in other clusters [22]. The mean Silhouette value for all documents is a measurement of the overall quality of a clustering solution with a specific number of clusters.

Since Silhouette values are based on distances, depending on the chosen distance measure different Silhouettes can be calculated. For instance, in both Figures 3 and 4 we use the complement of cosine similarity as distance measure, but in Figure 3 we consider text-based distances, while link-based distances are the input for Figure 4.

The reduction of the number of features in a vector space by application of LSI improves the performance of clustering and classification algorithms. In Section 4.1.1, we use Silhouette curves to contribute to the debate about the optimal number of LSI factors by investigating the relationship between number of factors, number of clusters, and clustering performance.

Stability.

The stability-based method of *Ben-Hur, Elisseeff*, and *Guyon* [3] allows to visually and quantitatively detect the most stable number of clusters from a stability diagram. The main idea is that the perceived structure should remain stable if only a subsample of objects is available, or if noise objects are added to the set. Multiple subsamples (e.g., 200) are randomly drawn from the data set, each comprising for instance 85% of objects. Then, a clustering algorithm subdivides each subsample into different numbers of clusters (e.g., 2 to 25 clusters). Next, the overlap between each pair of clustered subsamples is quantified by using, for example, the Jaccard coefficient (for a specific number of clusters).

Each number of clusters leads to one curve in the stability diagram (see Figure 5). The more a curve is to the right of the diagram, the higher the pairwise similarities between the clustered subsamples, and the more stable the clustering solutions with that specific number of clusters. A curve representing a certain number of clusters can be interpreted as how many percent of the subsample pairs (Y-axis) have a Jaccard value lower than or equal to the corresponding values on the X-axis. The number of clusters is chosen such that partitioning different subsamples leads to quite stable structures. In practice, a transition curve to the band of distributions on the left-hand side of the figure is selected.

3.4 Hybrid clustering

The requisite input for many clustering algorithms includes mutual distances between all objects (scientific publications here). These distances can be based on text, on citations, or on a combination of both information sources. The performance of clustering is even significantly improved by combinations of textual content with citations [13].

The idea of hybrid clustering is not new. For example, *He et al.* [10] have performed unsupervised spectral clustering of Web pages by combining textual information, hyperlink structure, and co-citation. Hyperlink structure was used as the dominant factor in the similarity metric and textual similarity was used to modulate the strength of each hyperlink. However, textual similarity between pages was neglected if both were not connected by a hyperlink. Integration with co-citation was achieved by a linear combination of co-citation and the weighted adjacency matrix of the graph. *Wang* and *Kitsuregawa* evaluated a contents-link coupled clustering algorithm for retrieved Web pages and studied the effect of out-links, in-links, specific terms, and their combination [24]. Results suggested that both links and contents are important for Web page clustering and that much better results are achieved with appropriate integration weights.

The remainder of this section describes a novel methodology for deeply combining text mining and bibliometrics by integrating text and citation information early in the mapping process, before application of the clustering algorithm.

For each data source, such as a normalized *term-by-document* matrix A or a normalized *cited_references-by-document* matrix B, square distance matrices D_t and D_{bc} can be constructed as follows:

$$\begin{aligned} D_t &= O_N - A^T \cdot A \\ D_{bc} &= O_N - B^T \cdot B, \end{aligned} \tag{1}$$

with N the number of documents and O_N a square matrix of dimensionality N with all ones. 'bc' refers to bibliographic coupling.

For the hybrid clustering of bioinformatics we used Fisher's inverse chi-square method to integrate textual similarity and citation information (bibliographic coupling). In the next Section we first briefly describe weighted linear combination of distance matrices.

3.4.1 Weighted linear combination

The distance matrices D_t and D_{bc} can be combined into an integrated distance matrix D_i by a weighted linear combination (linco) as follows:

$$D_i = \alpha \cdot D_t + (1 - \alpha) \cdot D_{bc} \tag{2}$$

The resulting D_i can then be used in clustering or classification algorithms. A comparable methodology was described as the toric k-means algorithm by *Modha* and *Spangler* [18]. Although being an attractive, easy, and reasonably scalable integration method, caution should be taken as a linear combination might neglect different distributional characteristics of various data sources [13] (p. 122–124).

The use of *Salton*'s cosine measure in both text and citation worlds leads to the same interval (range) of possible distances, but the actual distance distributions differ. The discrepancy in distributional characteristics can turn out even more severe when other information sources are considered. Different data matrices (such as *term-by-document* and *indicator-by-document*) may also require a different choice of distance metric. Differences in corresponding distributions might lead to an unequal or unfair contribution of both data sources in the ultimate integrated data, and might thus yield suboptimal results by implicitly favoring text over bibliometric information or vice versa. Spurious and strong (dis)similarities might obliterate good relationships established by the other data source.

3.4.2 Fisher's inverse chi-square method

As a plain linear combination might not be optimal for integrating textual and bibliometric information, we developed a methodology based on Fisher's inverse chi-square method. Fisher's inverse chi-square is an omnibus statistic from statistical meta-analysis to combine p-values from

multiple sources [11]. In contrast to the weighted linear combination procedure, this method can handle distances stemming from different metrics with different distributional characteristics and avoids domination of any information source.

Figure 2 illustrates the concept of distance integration by Fisher's inverse chi-square method. All text-based and link-based document distances in D_t and D_{bc}, as described above, are transformed to p-values with respect to the cumulative distribution function of distances for randomized data. This randomization is a necessary condition for having valid p-values. In our setting, a p-value means the probability that the similarity between two documents could be at least as high just by chance.

The randomized data sets can be constructed in several ways. The randomization should be as complete as possible, but should obey some rules that apply to the nature of the data. Blind randomizations might destroy important properties of human language. We considered different randomization schemes and finally opted for the somewhat conservative randomization which maintains the relative importance between terms by keeping the inverse document frequency for each term from the real data intact. Hence, term occurrences are randomly shuffled between documents, but the average characteristic document frequencies per term are preserved.

If the p-values for the textual data (p_1) and for link data (p_2) are calculated, an integrated statistic p_i can be computed as

$$p_i = -2 \cdot log(p_1^{\lambda} \cdot p_2^{1-\lambda}), \tag{3}$$

with $0 < \lambda < 1$ the integration weight determining the relative quality of both data sources and their contribution to the ultimate incorporated data. If the null hypothesis is true (i.e., in the case of randomized data), the distribution of $(p_1^{\lambda} \cdot p_2^{1-\lambda})$ is uniform and the integrated statistic has a chi-square distribution with 4 degrees of freedom [11]. The integrated p-value, p_i, is the new integrated document distance that can be used in clustering or classification algorithms.

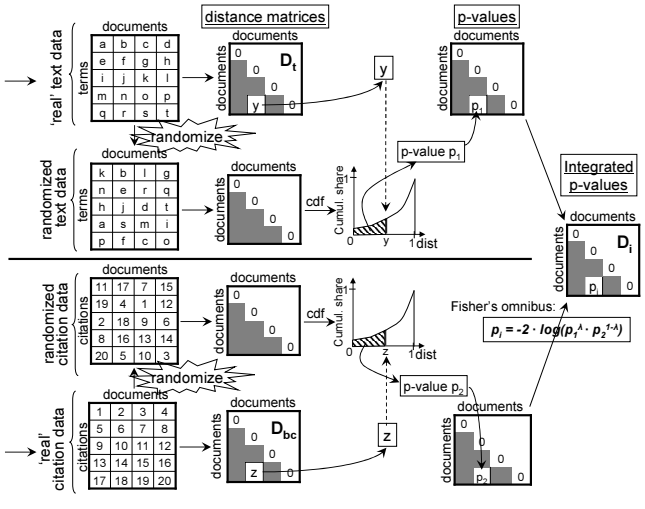


Figure 2: Distance integration by using Fisher's inverse chi-square method. The ultimate matrix with integrated p-values can be used for clustering.

Fisher's inverse chi-square method can also be applied if SVD is used as a pre-processing step for either the textual data (LSI), either for the citation-based component, or for both. The random document vectors are then first projected in the same space of reduced dimensionality before calculating the distribution of document similarities.

For more details about Fisher's inverse chi-square method for hybrid clustering, including a comparison with linear combination, we refer to [13]. In that publication we also demonstrate the performance on other data sets, such as a collection of full text documents about library and information science. We discuss how to estimate the integration weight λ, and present a rank-preserving modification to the original formula for BC and the superposition of a Gaussian noise factor in order to tackle a problem of discontinuity.

3.5 Dynamic hybrid clustering

Temporal analysis of text data has already been pursued by other authors. *Mei* and *Zhai* [17] introduced Temporal Text Mining to reveal evolutionary theme patterns in collections of news articles and scientific literature. They used probabilistic methods to generate and observe word clusters (themes) for different time periods. Kullback-Leibler divergence was used to discover coherent themes over time and Hidden Markov Models were applied to analyze the life cycle of each theme. *Griffiths* and *Steyvers* [8] have applied Latent Dirichlet Allocation and they explored basic temporal dynamics to identify hot topics in a collection of PNAS abstracts.

3.5.1 Matching & tracking clusters through time

Our strategy for dynamic clustering, namely by matching and tracking clusters through time, is demonstrated in Figure 7(a). Each horizontal level represents one period, indicated by the label of the leftmost circle and with a different gray level. Node size represents number of publications and for each cluster the best TF-IDF term is shown.

In each period, a separate hybrid clustering is performed (see Section 3.4) and the optimal number of clusters is again determined by observing the dendrogram, Silhouette curves, and *Ben-Hur* stability plot (see Section 3.3). Next, a complete graph is constructed with all cluster *centroids* from each period as nodes, and with as edge weights their mutual cosine similarities, calculated in the 10 dimensional latent semantic space.

3.5.2 Chains of clusters

Next, a two-step approach is followed in order to form *'cluster chains'*. First, only edge weights with similarity larger than a threshold $T1$ are retained. Secondly, clusters that fall below this threshold are yet allowed to join an existing chain if their similarity to *each* member is larger than threshold $T2$. In Figure 7(a), such clusters are depicted as a diamond instead of a circle.

3.5.3 Term networks

For visualization, we determined for each cluster chain the best words or phrases according to mean TF-IDF weights. In a term network (see Figure 7(b) for an example), each cluster chain has its own 'central node', represented by a diamond, which also indicates the number of publications. Each central node points to the best 10 keywords for the chain. When a keyword is among the best 10 for more than

one chain, it is only repeated once but connected to all corresponding cluster chain nodes. The gray level and thickness of an arc reflect the importance of a word for a cluster chain. Two terms are connected if both co-occur in one or more papers of the same cluster chain; the more co-occurrences, the closer the terms. Pajek was used for visualization [2].

A 'dynamic' term network allows the observation of shifts in vocabulary and focus of a specific (sub-)field of interest over time (see Figure 10). A central node is annotated with an indication of the period (such as '1991–1998'), with the period number, and with the number of publications.

4. HYBRID CLUSTERING RESULTS

4.1 Optimal number of clusters

4.1.1 Silhouette curves

The optimal combination of number of clusters and number of LSI factors depends on the document collection at hand and on the objectives of the study. In order to investigate the relationship between number of factors, number of clusters, and clustering performance, Figure 3 presents the mean Silhouette coefficient for 2 to 50 clusters, for different numbers of factors as well as for the standard Vector Space Model (no LSI). It is important to note that for the sake of comparability each distinct clustering was evaluated with Silhouette values calculated from the original term-by-document matrix A on which LSI had not been applied.

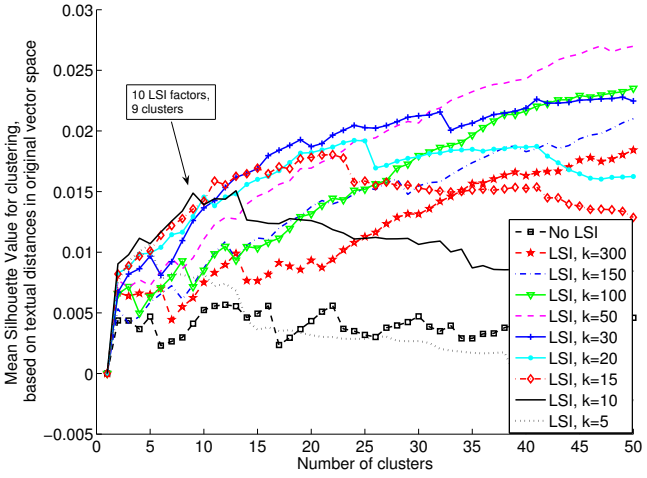


Figure 3: Silhouette curves with mean Silhouette coefficient for text-based clustering solutions of 2 up to 50 clusters, for the original term-by-document matrix ('No LSI') and for derived latent semantic indices ('LSI') with different numbers of factors k. The arrow indicates a local maximum for 9 clusters and 10 LSI factors.

Figure 3 demonstrates that, in general, the clustering performance is higher for a lower number of LSI factors (k). Nevertheless, the performance quickly drops if the number of clusters is higher than the number of factors. An explanation might be that it is a harder task to discern a certain number of clusters encoded in a lesser amount of dimensions. Hence, as a heuristic, it certainly seems advisable to use a number of factors at least as high as the desired number of clusters.

When looking for a coarse-grained clustering solution, a very modest number of factors, e.g., $k = 10$, might provide a local maximum in clustering performance. Indeed, Figure 3 shows that for any number of clusters less than 10, 10 LSI factors provide the best clustering performance. This observation is supported by a recent study by *Kontosthatis* and *Pottenger* [16]. They have shown in a retrieval setting that a small, fixed dimensionality reduction parameter ($k = 10$) can be used to capture the term relationship information in a corpus. A low dimensionality has also direct advantages in terms of storage needs and processing time.

Next, for more than 10 clusters, 15 factors take the lead, whereas 30 factors do best for finer-grained clustering solutions with more than 15 clusters. Again, from 31 clusters onwards, the next smallest number of factors in line, 50, is the winning number.

Although being all positive, the overall mean Silhouette values in Figure 3 each seem low, hinting at groups of documents that are not clearly separable according to the original classification of *Rousseeuw* [22]. This is probably due to the very high dimensionality of the original vector space in which the Silhouette values are computed (for comparability, as mentioned above), in contrast to the low-dimensional problems discussed in [22]. On the other hand, the nature of natural language usage might be of influence. When dealing with documents in comparable subject areas, the amount of overlapping words between different papers is, of course, considerable.

For the bioinformatics document set, 10 LSI factors and 9 clusters seem to be a very good combination (cf. the indicated local maximum in Figure 3). One might argue that other solutions with more clusters and more factors provide higher Silhouette values, but we are rather looking for a local maximum since we are not interested in 50 clusters within the bioinformatics field. Additional evidence is given by the Silhouette curve for link-based clustering using bibliographic coupling, which also shows a clear local maximum at 9 clusters (see Figure 4).

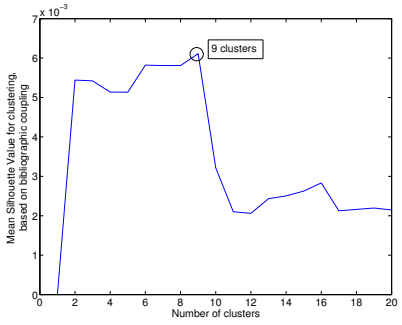


Figure 4: Silhouette curve with mean Silhouette coefficient for bibliographic coupling clustering.

4.1.2 Stability

Even more evidence for our 9 clusters within the bioinformatics field is provided by the stability diagram (see Section 3.3). The diagram of Figure 5 shows, for 2 up to 25 clusters, the cumulative distribution of pairwise Jaccard simi-

larities, between 200 pairs of clustered random subsamples, each comprising 6291 bioinformatics publications (sampling ratio of 85%).

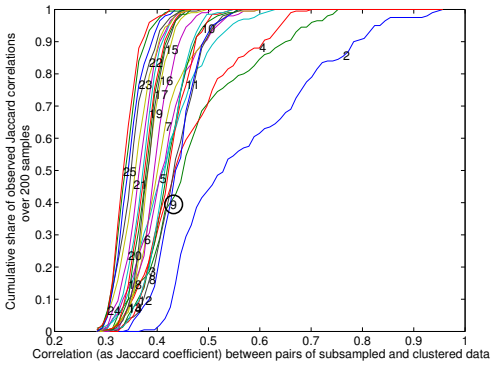


Figure 5: Stability diagram.

Although the most stable solution is obtained for partitioning the bioinformatics papers in two clusters, we are looking for a finer-grained clustering. Nine clusters prove much more stable than 5, 6 or 7 clusters, and compete with solutions of 3 and 4 clusters. 8 clusters are almost as stable as 9 clusters.

4.1.3 *Dendrogram*

Figure 6 depicts the dendrogram cut off at 9 clusters on the left-hand side, which seemed a good cut-off point. For each of 9 clusters, the number of publications and the best mean TF-IDF term or phrase are shown. These automatically determined labels already give a quite good impression of the contents of the clusters.

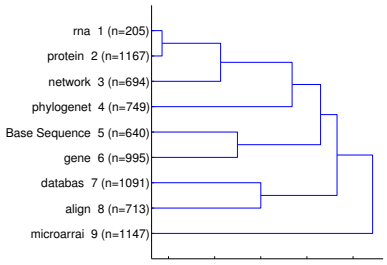


Figure 6: Dendrogram, cut off at 9 clusters on the left-hand side.

4.1.4 *Concluding remark*

Although the number of clusters remains a difficult to define parameter, our experience is that the different strategies often agree on a certain local maximum of clustering performance. For the bioinformatics field the consensus was 9 clusters.

4.2 Cluster characterization

After observing the contents of all clusters in detail we were able to propose a name for each cluster as given in Table 1. The cluster size is indicated as well, next to the characterization by most salient author keyword and by best TF-IDF MeSH term. For MeSH terms, the *TF* factor was either 1 or 2, for *minor* and *major* MeSH descriptors, respectively. With 205 publications, cluster 1, labeled *RNA structure prediction*, is the smallest one; all other clusters have more than 600 and less than 1200 papers. We determined 'representative' publications by using four different methods that rank the papers in each cluster according to different criteria of importance. Table 1 lists for each cluster the paper on top of each ranking: *(1)* the *medoid*, which is the paper that is most similar to the mean cluster profile (the *centroid*), *(2,3)* the best *authority* and best *hub* determined by the HITS algorithm, and *(4)* the paper with highest PageRank.

5. DYNAMIC CLUSTERING RESULTS

Figure 7(a) visualizes the dynamic clustering strategy, which is explained in Section 3.5. The two thresholds $T1 = 0.95$ and $T2 = 0.8$ were determined by observing Figure 8. Similarities above $T1$ are very high, but nevertheless do occur more often than somewhat less pronounced matches. We consider these very high similarities as corresponding to true cluster matches. Somewhat surprisingly, most cluster chains were well established after the first step with stern requirement ($T1$). Additional members, added by $T2$, are indicated with a diamond instead of a circle in Figure 7(a).

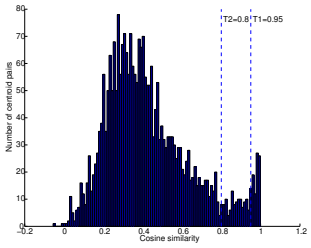


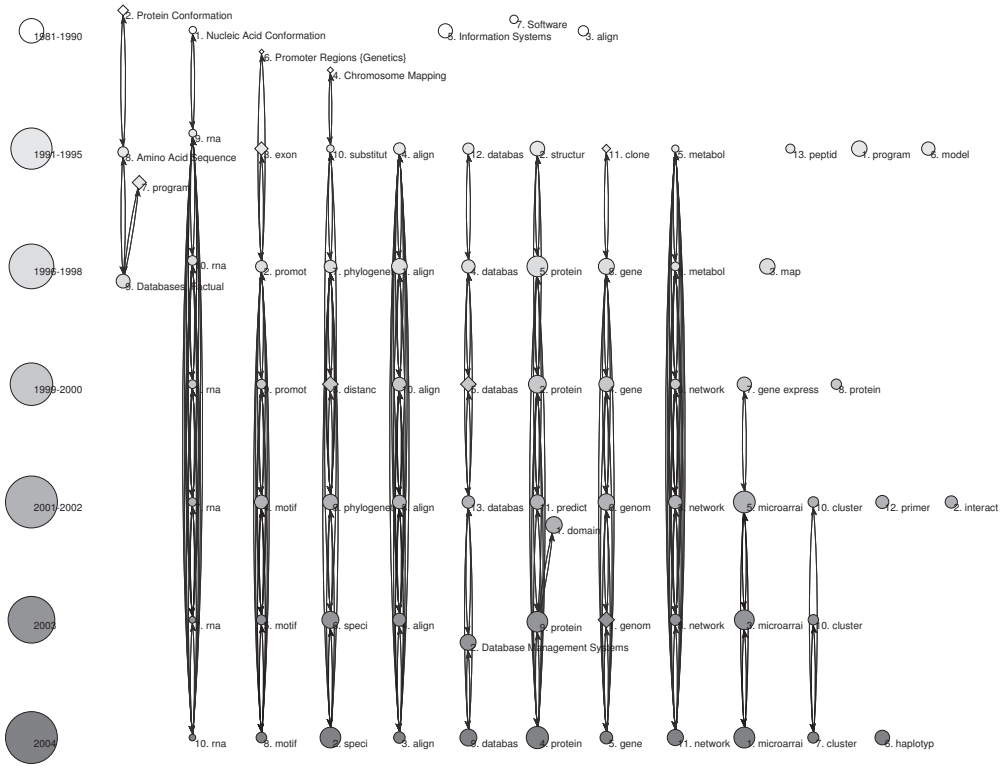
Figure 8: Histogram of mutual similarities between all cluster centroids of Figure 7(a). The demarcation of strong ($T1$) and less strong ($T2$) cluster matches was defined visually.

5.1 Cluster chains

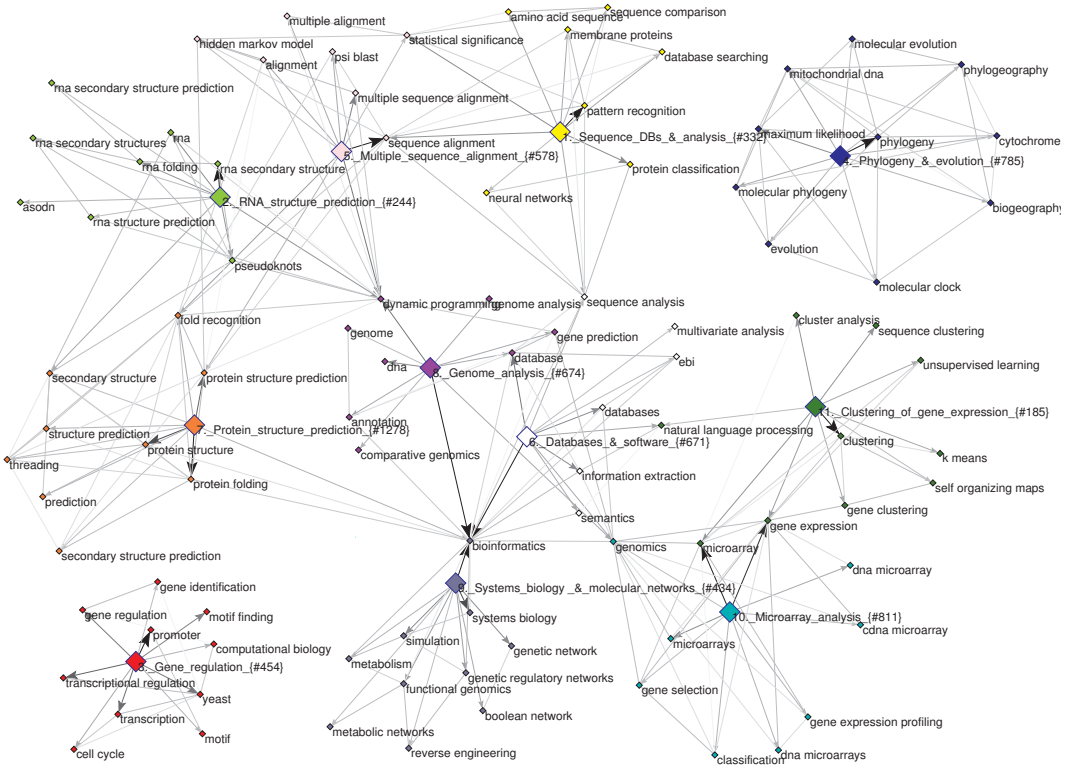
Whereas the 'static' hybrid clustering algorithm came up with 9 clusters, figure 7(a) suggests that in total 11 *cluster chains* could be distinguished by the dynamic procedure, 3 of which contain publications from all seven periods between 1981 and 2004. Five chains emerged in 1991 and were still present in 2004. The *'microarrai'* chain (10th from left to right) appeared in 1999–2000 and the *'cluster'* chain (#11) one period later (2001–2002). These two chains together approximately constitute the *'Microarray analysis'* cluster of Table 1. The chain on the left-hand side lasted from the first until the third period. Besides these groups of documents that are connected in cluster chains, some others are not connected to any chain. By disregarding these clusters that could not be linked to any other cluster in another period, the dynamic methodology of tracking clusters through time can be considered less *'hard'* than the standard hierarchical clustering algorithm in the sense that not all publications need be attributed to at least one chain.

Table 1: For each of 9 clusters: (1) the publication with largest cosine similarity to the mean cluster profile (*medoid* paper); (2) the best authority and (3) the best hub paper detected by the HITS algorithm; and (4) the paper with highest PageRank according to Google's algorithm. Cluster names and sizes are indicated as well, next to the characterization by most salient author keyword and by best TF-IDF MeSH term. Only first authors are indicated.

Cluster 1. RNA structure prediction (n=205; rna; Nucleic Acid Conformation)
1) Major. Computational methods for RNA structure determination. *Curr Opin Struc Biol* 11(3):282-286, 2001.
2) Mathews. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 288(5):911-940, 1999.
3) Zuker. Calculating nucleic acid secondary structure. *Curr Opin Struc Biol* 10(3):303-310, 2000.
4) Zuker. Optimal Computer Folding of Large RNA Sequences Using Thermodynamics and Auxiliary Information. *Nucleic Acids Res* 9(1):133-148, 1981.
Cluster 2. Protein structure prediction (n=1167; protein; Proteins/chemistry)
1) Di Francesco. FORESST: fold recognition from secondary structure predictions of proteins. *Bioinformatics* 15(2):131-140, 1999.
2) Murzin. Scop - A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *J Mol Biol* 247(4):536-540, 1995.
3) Eisenhaber. Protein-Structure Prediction - Recognition of Primary, Secondary, and Tertiary Structural Features from Amino-Acid-Sequence. *Crit Rev Biochem Mol* 30(1):1-94, 1995.
4) Chothia. The Relation Between the Divergence of Sequence and Structure in Proteins. *Embo J* 5(4):823-826, 1986.
Cluster 3. Systems biology & molecular networks (n=694; network; Models, Biological)
1) Xiong. Network-based regulatory pathways analysis. *Bioinformatics* 20(13):2056-2066, 2004.
2) Jeong. The large-scale organization of metabolic networks. *Nature* 407(6804):651-654, 2000.
3) Xia. Analyzing cellular biochemistry in terms of molecular networks. *Annu Rev Biochem* 73:1051-1087, 2004.
4) Karp. EcoCyc: Encyclopedia of Escherichia coli genes and metabolism. *Nucleic Acids Res* 26(1):50-53, 1998.
Cluster 4. Phylogeny & evolution (n=749; phylogenet; Phylogeny)
1) Negrisolo. Morphological convergence characterizes the evolution of Xanthophyceae (Heterokontophyta): evidence from nuclear SSU rDNA and plastidial rbcL genes. *Mol Phylogenet Evol* 33(1):156-170, 2004.
2) Posada. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14(9):817-818, 1998.
3) Delsuc. Molecular systematics of armadillos (Xenarthra, Dasypodidae): contribution of maximum likelihood and Bayesian analyses of mitochondrial and nuclear genes. *Mol Phylogenet Evol* 28(2):261-275, 2003.
4) Saitou. The Neighbor-Joining Method - A New Method for Reconstructing Phylogenetic Trees. *Mol Biol Evol* 4(4):406-425, 1987.
Cluster 5. Genome sequencing & assembly (n=640; base sequenc; Base Sequence)
1) Barber. Sequenceeditingaligner - A Multiple Sequence Editor and Aligner. *Genet Anal-Biomol Eng* 7(2):39-45, 1990.
2) SantaLucia. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *P Natl Acad Sci USA* 95(4):1460-1465, 1998.
3) Kaderali. Selecting signature oligonucleotides to identify organisms using DNA arrays. *Bioinformatics* 18(10):1340-1349, 2002.
4) Wilbur. Rapid Similarity Searches of Nucleic-Acid and Protein Data Banks. *P Natl Acad Sci USA-BIOL SCI* 80(3):726-730, 1983.
Cluster 6. Gene/promoter/motif prediction (n=995; gene; Sequence Analysis, DNA/methods)
1) Park. Comparing expression profiles of genes with similar promoter regions. *Bioinformatics* 18(12):1576-1584, 2002.
2) Burge. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 268(1):78-94, 1997.
3) Mathe. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res* 30(19):4103-4117, 2002.
4) Uberbacher. Locating Protein-Coding Regions in Human DNA-Sequences by A Multiple Sensor Neural Network Approach. *P Natl Acad Sci USA* 88(24):11261-11265, 1991.
Cluster 7. Molecular DBs & annotation platforms (n=1091; databas; Databases, Factual)
1) Andrade. Automated genome sequence analysis and annotation. *Bioinformatics* 15(5):391-412, 1999.
2) Bairoch. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 28(1):45-48, 2000.
3) Kriventseva. Clustering and analysis of protein families. *Curr Opin Struc Biol* 11(3):334-339, 2001.
4) Henikoff. Automated Assembly of Protein Blocks for Database Searching. *Nucleic Acids Res* 19(23):6565-6572, 1991.
Cluster 8. Multiple sequence alignment (n=713; align; Sequence Alignment/methods)
1) Jaroszewski. Improving the quality of twilight-zone alignments. *PROTEIN SCI* 9(8):1487-1496, 2000.
2) Altschul. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389-3402, 1997.
3) Gotoh. Multiple sequence alignment: Algorithms and applications. *ADV BIOPHYS* 36:159-206, 1999.
4) Fitch. Optimal Sequence Alignments. *P Natl Acad Sci USA-Biological Sciences* 80(5):1382-1386, 1983.
Cluster 9. Microarray analysis (n=1147; microarrai; Oligonucleotide Array Sequence Analysis/methods)
1) Tsai. An evolutionary approach for gene expression patterns. *IEEE T Inf Technol B* 8(2):69-78, 2004.
2) Eisen. Cluster analysis and display of genome-wide expression patterns. *P Natl Acad Sci USA* 95(25):14863-14868, 1998.
3) Hackl. Analysis of DNA microarray data. *Curr Top Med Chem* 4(13):1357-1370, 2004.
4) Schena. Quantitative Monitoring of Gene-Expression Patterns with A Complementary-DNA Microarray. *Science* 270(5235):467-470, 1995.



(a)



(b)

Figure 7: Dynamic hybrid clustering. (a). Matching and tracking clusters through time. Each horizontal level represents one period, indicated by the label of the leftmost circle and with a different gray level. Node size represents number of publications and for each cluster the best TF-IDF term is shown. (b). Term networks for the *cluster chains* with the best 10 author keywords according to mean TF-IDF scores.

Subsets of documents that do not clearly belong to any of the chains can be neglected. Hence, the chains that have been found by the dynamic clustering procedure might be more accurate than the clusters found with the standard algorithm, but this should be further assessed in subsequent experiments. If for a certain period a non-optimal number of clusters would be chosen, the strategy of tracking and matching of clusters through time can compensate for it by joining more clusters of that period to the same cluster chain. Likewise, a cluster chain might also split up into different branches, when, for example, two centroids of a later period are both linked to the same one of a previous period and both develop further in dissociated chains. In our data set, such dissociation is not observable, but the joining of two centroids of the same period in one chain is. If a line of research would be discontinued in a certain period, but be resumed again in a later one, this would also be detected and the resulting chain would just bridge the period with no activity in that area. A drawback, however, is that some clusters can still be overlooked by application of the visually defined, simple similarity thresholds. Improvement for the dynamic methodology might be obtained by using more complex rules for the forming of the chains of clusters.

5.2 Term networks

Figure 7(b) depicts the cognitive structure of bioinformatics by showing, for each of 11 cluster chains, the term networks with the best 10 author keywords. The central node of each term network reveals the chain number, the chain name, and the number of publications in the chain.

5.3 Dynamics

By visualizing the relative activity in the different chains, Figure 9 provides a view on how much attention the bioinformatics community has devoted to the different subfields through time. The share (in %) of the yearly publication

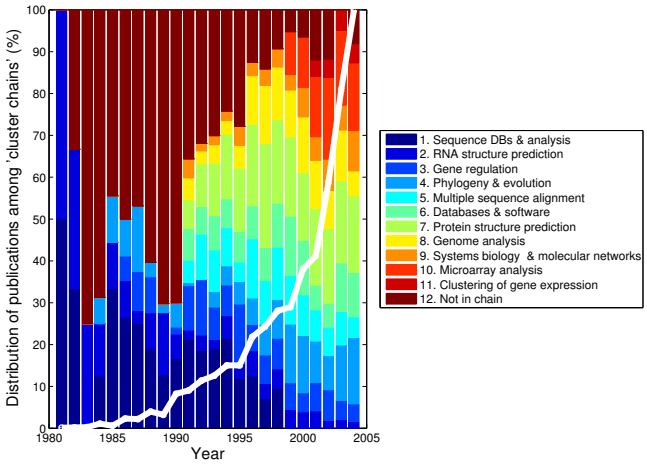


Figure 9: Distribution of the total yearly publication output among cluster chains. The white line indicates the yearly number of publications, relative to the number in 2004 (1455). 'Chain' 12 represents all publications that are *not* connected to any chain.

output that belongs to each cluster chain is shown with a different color. The white line depicts the yearly number of bioinformatics publications relative to the number of publications that were published in the year 2004 (1455). This way of visualizing demonstrates the relative growing and fading of different topics in bioinformatics. It is clear that the share of publications *Not in chain* (#12) diminishes mostly with respect to previous years. This is an indication of the bioinformatics field starting to form crisp lines of research, especially after the year 1990. An upward trend in relative number of publications can be ascribed to the chains *Microarray analysis* (#10), *Systems Biology & molecular networks* (#9) and *Phylogeny & evolution* (#4). The first two are recent subfields in which a lot of scientific research is being conducted today. Cluster chain 4, *Phylogeny & evolution*, actually represents a relatively old research field, but new developments in bioinformatics made it regain a lot of attention since the start of the new millennium. Some of the cluster chains, e.g., *RNA structure prediction* (#2), represent 'older' subfields that are relatively almost fading away.

Figure 10 shows a *dynamic* term network for the *Systems Biology & molecular networks* cluster chain (#9), with for each period the best 10 author keywords according to mean TF-IDF scores. Computational Systems Biology studies biological systems at various scales, the building blocks and how these form networks of relationships. Dynamic quantitative models are built based on properties of the components, and even enable predictions.

The central node in the upper part of Figure 10 (*1996-1998_2_metabol_#42*) corresponds to the period 1996–1998, which accounts for 42 papers. It is a bit isolated in the sense that none of its terms are also among the best for another chain, and no term has co-occurred with one of the salient terms of another period. Looking in a clockwise manner, temporal keywords of successive periods are illustrated.

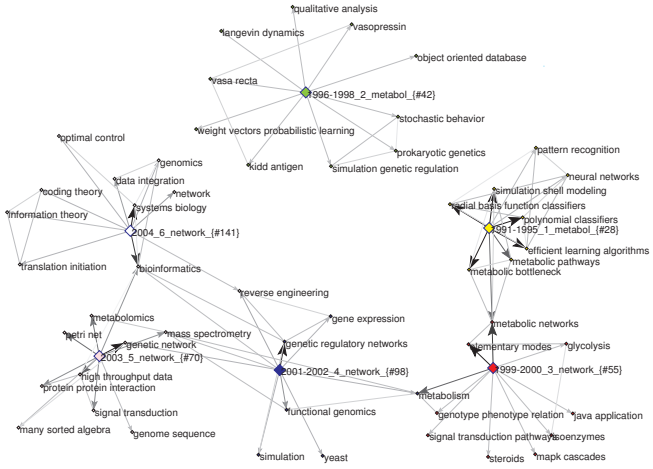


Figure 10: Dynamic term network for the *Systems Biology & molecular networks* cluster chain (#9).

6. CONCLUDING REMARKS

The demarcation of the interdisciplinary field of bioinformatics was achieved by bibliometric retrieval. Seven consecutive periods containing approximately the same number of publications were defined.

We demonstrated the hybrid clustering method based on Fisher's inverse chi-square, which was revealed in previous

research as a promising method for integrating textual content and citations. It significantly outperforms text-only and link-only methods, as well as other integration schemes.

We investigated the relationship between number of Latent Semantic Indexing factors, number of clusters, and clustering performance. In general, the quality of clustering proved significantly higher for a smaller number of LSI factors. In our data set, a very modest number of factors (e.g., 10) delivers local maxima in clustering performance, on condition that there are no fewer LSI factors than the desired number of clusters.

A combined strategy for determination of the optimal number of clusters, comprising distance-based and stability-based methods, suggested nine subdisciplines. For each cluster we provided the medoid and other representative publications according to HITS and PageRank applied to the citation graph.

Next, a methodology was developed for dynamic clustering. The same hybrid clustering algorithm was applied multiple times, but each time restricted to publications in one of the defined periods. Eleven *cluster chains* could be identified by matching and tracking clusters through time. Their concept networks and evolution were analyzed.

To conclude, the hybrid clustering algorithm exploiting information from both the text and citation worlds, possibly complemented with the dynamic strategy of tracking clusters through time, provide powerful tools to unravel the cognitive structure of scientific or technological fields, to cast eyes upon the evolution of existing subdisciplines, and to aid detection of emerging or converging clusters.

7. ACKNOWLEDGEMENTS

Research supported by Research Council KUL: GOA AM-BioRICS, CoE EF/05/007 SymBioSys, several PhD/postdoc & fellow grants. Flemish Government: Steunpunt O&O Indicatoren; FWO: PhD/postdoc grants, projects G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, GBOU-McKnow-E (Knowledge management algorithms), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; SBO-BioFrame; Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011) ; EU-RTD: ERNSI: European Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain, FP6-STREP Strokemap.

8. REFERENCES

[1] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., 1999.

[2] V. Batagelj and A. Mrvar. Pajek - analysis and visualization of large networks. *Graph Drawing*, 2265:477–478, 2002.

[3] A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*, pages 6–17, 2002.

[4] M. W. Berry, S. T. Dumais, and G. W. Obrien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595, 1995.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[7] W. Glänzel, F. Janssens, and B. Thijs. A comparative analysis of publication activity and citation impact based on the core literature in bioinformatics. In *Proc. 11th Intl. Conf. of the ISSI, Madrid, Spain*, 2007.

[8] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101:5228–5235, 2004.

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

[10] X. He, C. H. Q. Ding, H. Zha, and H. D. Simon. Automatic topic identification using webpage clustering. In *Proc. of the 2001 IEEE intl. conf. on Data Mining*, pages 195–202, Washington, DC, USA, 2001. IEEE Computer Society.

[11] L. V. Hedges and I. Olkin. *Statistical Methods for Meta-Analysis*. Academic Press, 1985.

[12] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[13] F. Janssens. *Clustering of scientific fields by integrating text mining and bibliometrics*. Ph.D. thesis, Faculty of Engineering, Katholieke Universiteit Leuven, Belgium, http://hdl.handle.net/1979/847, 2007.

[14] M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25, 1963.

[15] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[16] A. Kontostathis and W. M. Pottenger. Essential Dimensions of Latent Semantic Indexing (EDLSI). In *Proc. 40th Annual Hawaii Intl. Conf. on System Sciences (CD-ROM)*, 2007.

[17] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *KDD '05*, pages 198–207, New York, NY, USA, 2005. ACM Press.

[18] D. S. Modha and W. S. Spangler. Clustering hypertext with applications to web searching. In *Proc. of the 11th ACM Conf. on Hypertext and Hypermedia*, pages 143–152, New York, 2000. ACM Press.

[19] C. A. Ouzounis and A. Valencia. Early bioinformatics: the birth of a discipline - a personal view. *Bioinformatics*, 19(17):2176–2190, 2003.

[20] S. K. Patra and S. Mishra. Bibliometric study of bioinformatics literature. *Scientometrics*, 67(3):477–489, 2006.

[21] C. Perez-Iratxeta, M. A. Andrade-Navarro, and J. D. Wren. Evolving research trends in bioinformatics. *Briefings in Bioinformatics*, 2006.

[22] P. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, 1987.

[23] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986.

[24] Y. Wang and M. Kitsuregawa. Evaluating contents-link coupled web page clustering for web search results. In *Proc. 11th intl. conf. on Information and Knowledge Management*, pages 499–506, New York, NY, USA, 2002. ACM Press.