A Complex Bio-networks of the Function Profile of Genes

Charles C. H. Liu[1,3], I-Jen Chiang[2,3,⋆], Jau-Min Wong[3], Ginni Hsiang-Chun Tsai[3], and Tsau Young ('T. Y.') Lin[4]

[1] Department of Surgery, Cathay Medical Center, Taipei, Taiwan 106
chliu@ntu.edu.tw
[2] Graduate Institute of Medical Informatic, Taipei Medical University,Taipei, Taiwan 110
ijchiang@tmu.edu.tw
[3] Graduate Institue of Medical Engineering, National Taipei University, Taipei, Taiwan 100
[4] Department of Computer Science, San Jose State University, San Jose, CA 95192-0249
tylin@cs.sjsu.edu

Abstract. This paper presents a novel model of concept representation using a multilevel geometric structure, which is called *Latent Semantic Networks*. Given a set of documents, the associations among frequently co-occurring terms in any of the documents define naturally a geometric complex, which can then be decomposed into connected components at various levels.

This hierarchical model of knowledge representation was validated in the functional profiling of genes. Our approach excelled the traditional approach of vector-based document clustering by the geometrical forms of frequent itemsets generated by the association rules. The biological profiling of genes were a complex of concepts, which could be decomposed into primitive concepts, based on which the relevant literature could be clustered in adequate "resolution" of contexts. The hierarchical representation could be validated with tree-based biomedical ontological frameworks, which had been applied for years, and been recently enriched by the online availability of *Unified Medical Language System (UMLS)* and *Gene Ontology (GO)*.

Demonstration of the model and the clustering would be performed on the relevant *GeneRIF (References into Function)* document set of NOD2 gene. Our geometrical model is suitable for representation of bio-logical information, where hierarchical concepts in different complexity could be explored interactively according to the context of application and the various needs of the researchers. An online clustering search engine for use on general purpose and for biomedical use, managing the search results from Google or from PubMed, are constructed based on the methodology (http://ginni.bme.ntu.edu.tw). The hierarchical presentation of clustering results and the interactive graphical display of the contents of each cluster shows the merits of our approach.

1 Introduction

One of the urgent need of bioinformatics in the post-genomic era is to find "biological themes" or "topics" between genes or gene products, in order to "drink from the fire hose" from vast amounts of literature and ex-periment results.

⋆ Corresponding author.

C. Priami et al. (Eds.): Trans. on Comput. Syst. Biol. V, LNBI 4070, pp. 9–24, 2006.

One approach of theme finding is to derive knowledge directly without translation by another knowledge source, e.g. a vocabulary system. One of the early successful approaches is direct mining from the source literature. The relationships between genes are constructed by probabilistic modes, such as Bayesian Networks. The most clinically yielding is the PubGene project [4]. However, the interpretation of the results is often qualitative, selectively on some local findings in large graph models. The lack of overall picture is partly due to the exploration of individual genes without preliminary grouping of some closely correlated genes. The result relied on the quality of documents collected as "relevant" to the target genes [8].

Subsequent researches to find "molecular pathway" in raw documents is vigorous use of natural language processing techniques. One of the efforts with a long history of literature mining in other medical domain is the GENIE project, evolved from MEDLEE works [2]. Finely tuned rule-based term tagging and processing improve the efficiency, but the rule sets or knowledge sources they constructed cannot be reused by other applications or be validated by others. Besides, the system is too large for personal document browsing.

The other approaches use external knowledge system, such as keyword hierarchy, to group the raw gene information to more biologically under-standable "themes". The early works are well reviewed by Shatkay in the analysis of microarray data [8]. MedMesh is more recent work addressing on the MeSH systems (Medical Subject Heading) of UMLS (Unified Medical Language System), but much raw document processing is used and the approach was relatively in a "black box" [6]. After the advent of Gene Ontology (GO) system, more tools were developed to apply the ontological framework to impose domain knowledge on analysis of raw data, which were listed under the section of "GO tools" in the official site of the GO Consortium [1].

From the medical point of view, current application of MeSH or GO is still in a very primitive developing stage. One of the main reason lies on the nature of tree-based ontological system. For example, GO divides the functional profiles into three branches from the root – the function do-main, the process domain, and the anatomical domain. The first two do-mains are closely associated in many applications. The third domain is also dependent on the first two "function" domains. In addition, the amount of annotations of genes to the three domains is also unbalanced.

Our research addresses on the limitation of functional analysis of genes by the traditional approaches, and proposed a new geometric model. In what follows, we start by reviewing related work on the models of the relationships between gene and gene products clustering in section 2. The concepts and definitions of *latent semantic networks* based on geometric forms for the frequent itemsets generated by association rules are given in section 3. The clustering results for clustering of the functioning profile of a gene are described in Section 4 and Section 5; followed by the conclusion.

2 Related Work

Detecting knowledge based on the co-occurrence of terms or concepts is one of the basic mechanism of document clustering, and was initially proposed to cluster genes into biologically meaningful groups [4]. However, the characteristics of the "groups"

could not be explained by the co-occurrence alone. An approach of getting the biological "meaning" was by annotation with associated MeSH and GO terms, which were both tree-based. Our work approaches the "meaning" problem by proposing a new geometric model of clustering in order to more adequately present the network nature of the functioning profiles of genes.

After Girvan and Newman's work of "community structure" in social and biological networks [3], the nature of graph structure inherent in a co-occurrence network began to be explored. Wilkinson et al. [6] picked sets of genes correlated to user-selected keywords by partitioning the components of gene co-occurrence networks functionally correlated "communities". Wren et al. [9] studied the connections in the gene network to rank the "cohesiveness" of co-occurring genes, diseases, and chemical compounds.

The current published genetic analyses based on "community networks" were calculated based on geometrical measurement in the Euclidean space, which we considered is a fundamental limitation of statistical calculation in document or concept clustering. The clustering of distance measurements between sets of more primitive concepts to form higher hierarchy of concept groups is more applicable in topological spaces than in Euclidean spaces. We proposed a topologically based network more suitable for gene analysis.

Based on the network model, we constructed an online "clustering search" engine, which received PubMed and Google queries results, selected the significant concepts, and provided hierarchical and graphical views of the associations between concepts on the fly. The details could be explored by the readers according to their needs interactively. By the example of recently vigorously explored NOD2 gene, which is closely associated with the immunity and inflammatory responses of the inflammatory bowel disease (IBD), we demonstrated the practical applications of our model and clustering methodology in functional profiling of genes. Automatic generation of graphical relationships between concepts from the user specified scope also provoked new insights into the structure of knowledge and was beneficial for improvement of current tree-based vocabulary systems.

3 Geometric Representation of Concept

Term-term inter-relationships that are denoted by their co-occurred associations can automatically model and extract the concepts from a collection of documents. These concepts organize a multilevel and homogenous hierarchy called a *Latent Semantic Network*. The most natural way to represent a latent semantic network is expressed by using the geometric and topologic notations, which can capture the totality of thoughts expressed in this collection of documents; and a "simple component" (which is a r*-connected component*) of a level of hierarchy represents some concept inside this collection.

3.1 Combinatorial Geometry

Let us introduce and define some combinatorial topological concepts. The central idea is n-simplex.

Definition 1. *A* n*-simplex is a set of independent abstract vertices* $[v_0, \ldots, v_{n+1}]$.

Geometrically 0-simplex is a vertex, 1-simplex an edge (a vertex pair), 2-simplex a triangle, 3-simplex a tetrahedron. A n-simplex is the $n+1$ dimensional analog. This is the smallest convex set in a Euclidean space R^{n+1} containing $n+1$ points $v_0 \ldots$, v_{n+1} that do not lie in a hyperplane of dimension less than n. For example, there is the standard n-simplex

$$\delta^n = \{(t_0, t_1, \ldots t_{n+1}) \in R^{n+1} \mid \sum_i t_i = 1, t_i \geq 0\}$$

Definition 2. *A face of a n-simplex $[v_0, \ldots, v_{n+1}]$ is a r-simplex $[v_{j_0}, \ldots, v_{j_{r+1}}]$ whose vertices is a subset $\{ v_0, \ldots, v_{n+1} \}$ with cardinality $r+1$.*

Definition 3. *A complex is a finite set of simplices that satisfies the following two conditions:*

- *Any face of a simplex from a complex is also in this complex.*
- *The intersection of any two simplices from a complex is either empty or is a face for both of them.*

The vertices of the complex v_0, v_1, $\cdots$, v_n is the union of all vertices of those simplices. [7]

Definition 4. *A* hereditary n simplex*, or abbreviated to be* n-H-simplex *is a special complex of n dimensions that consists of one n-simplex and all its faces.*

Definition 5. *A (n, r)-skeleton (denoted by S_r^n) of n-complex is a n-complex whose k-faces($k \leq r$) are removed.*

Definition 6. *For any non-empty two simplices A, B are said to be* r-connected *if there exits a sequence of k-simplices $A = S_0, S_1, \ldots, S_m = B$ such that S_j and S_{j+1} has an h-common face for $j = 0, 1, 2, \ldots, m-1$; where $r \leq h \leq k \leq n$.*

Definition 7. *The maximal r-connected subcomplex is called a* r-connected component*. Note For a r-connected component implies there does not exist any r-connected component that is the superset of it.*

3.2 Simple Concept Geometric Structure

In our application each vertex is a key term, so a simplex defines a set of key terms in a collection of documents. Hence, we believe a simplex represents a primitive concept in the collection. For example, the 1-simplex [Wall, Street] represents a primitive concept in financial business. The 0-simplex [Network] might represent many different concepts, however, while it is combined with some other terms would denote latent semantic concepts, such as, these 1-simplices [Computer, Network], [Traffic, Network], [Neural, Network], [Comunication, Network], and so on, demonstrate distinct concepts and identify more obvious semantic than 0-simplex. Of course, the 1-simplex [Neural, Network] is not conspicuous than the 2-simplices [Artifical Neural Network] and [Biology, Neural, Network].

A collection of documents most likely consists of several distinct primitive concepts. Such a collection of primitive concepts is combinatorial a complex.

An idea (in the forms of complex of keywords) may consist of a lot of primitive concepts (in the form of simplices) that are embedded in a document collection. Some primitive concepts may share a common primitive concept, some may not. This situation may be captured by a combinatorial complex of key terms: An idea in the forms of a complex of keywords may consist of a lot of primitive concepts in the form of simplices. Some primitive concepts (simplices) may share a common concept (a common face), some may not.

Example 1. *In Figure 4, we have an idea that consist of twelve terms that organized in the forms of* 3*-complex. Two* $Simplex(a, b, c, d)$ *and* $Simplex(w, x, y, z)$ *are two maximal H-simplices with the highest rank* 3*. Considering* $(3, 1)$*-skeleton,* S^3_1*, by removing*

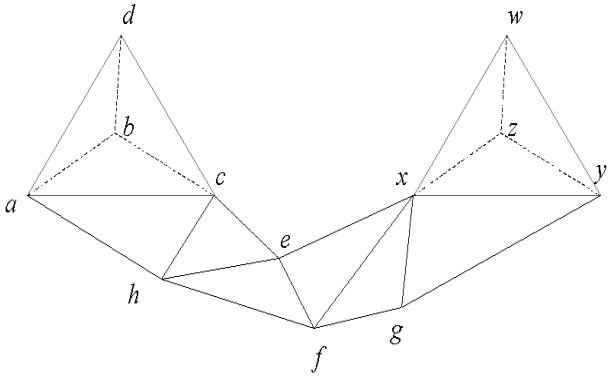

Fig. 1. A complex with twelve vetrics

all 0*-simplices, all the other simplices in it can be listed as follows.*

- Simplex(a, b, c, d) *and its ten subsimplices:*
  - Simplex(a, b, c)
  - Simplex(a, b, d)
  - Simplex(a, c, d)
  - Simplex(b, c, d)
  - Simplex(a, b)
  - Simplex(a, c)
  - Simplex(b, c)
  - Simplex(a, d)
  - Simplex(b, d)
  - Simplex(c, d)
- Simplex(a, c, h) *and its three subsimplices:*
  - Simplex(a, c)
  - Simplex(a, h)
  - Simplex(c, h)

- Simplex(c, h, e) *and its three subsimplices:*
  - Simplex(c, h)
  - Simplex(h, e)
  - Simplex(c, e)
- Simplex(e, h, f) *and its three subsimplices:*
  - Simplex(e, h)
  - Simplex(h, f)
  - Simplex(e, f)
- Simplex(e, f, x) *and its three subsimplices:*
  - Simplex(e, f)
  - Simplex(e, x)
  - Simplex(f, x)
- Simplex(f, g, x) *and its three subsimplices:*
  - Simplex(f, g)
  - Simplex(g, x)
  - Simplex(f, x)
- Simplex(g, x, y) *and its three subsimplices:*
  - Simplex(g, x)
  - Simplex(g, y)
  - Simplex(x, y)
- Simplex(w, x, y, z) *and its ten subsimplices:*
  - Simplex(w, x, y)
  - Simplex(w, x, z)
  - Simplex(w, y, z)
  - Simplex(x, y, z)
  - Simplex(w, x)
  - Simplex(w, y)
  - Simplex(w, z)
  - Simplex(x, y)
  - Simplex(x, z)
  - Simplex(y, z)

Simplex(a, c), Simplex(c, h), Simplex(h, e), Simplex(e, f), Simplex(f, x), Simplex (g, x), *and* Simplex(x, y) *are common faces that generate a connected path from* Simplex(a, b, c, d) *to* Simplex(w, x, y, z). *There exists a single maximal connected component. Furthermore, considering the* $(3, 2)$*-skeleton,* S_2^3*, by eliminating all* 0*-simplices and* 1*-simplices, all the remainder simplices of it are as follows.*

- Simplex(a, b, c, d) *and its four subsimplices:*
  - Simplex(a, b, c)
  - Simplex(a, b, d)
  - Simplex(a, c, d)
  - Simplex(b, c, d)
- Simplex(a, c, h)
- Simplex(c, h, e)
- Simplex(e, h, f)
- Simplex(e, f, x)

- $\mathrm{Simplex(f, g, x)}$
- $\mathrm{Simplex(g, x, y)}$
- $\mathrm{Simplex(w, x, y, z)}$ *and its four subsimplices:*
  - $\mathrm{Simplex(w, x, y)}$
  - $\mathrm{Simplex(w, x, z)}$
  - $\mathrm{Simplex(w, y, z)}$
  - $\mathrm{Simplex(x, y, z)}$

There does not exist any common faces between any two simplices, so that eight maximal connected components are in S_2^3. So does S_3^3, there are only two maximal connected components in it because the maximum rank of simplices in it is 3.

A maximal connected component of a skeleton represents a complex of association rules, i.e., a set of concepts. If a maximal connected component of a skeleton contains only one simplex, this component is said to organize a primitive concept.

Definition 8. *A maximal connected component is said to be* independent *if it is composed of a single simplex, i.e., there is no common face between two maximal connected components.*

3.3 Issues

From a collection of documents, a complex of association rules can be generated. A skeleton of a complex is closed, because all subcomplexes of a complex are also in the skeleton according to subsimplices in each composite simplex of a complex in a skeleton are also included in the simplex, which satisfies the *apriori* property. As seen in Example 1, all connected components in S_k^n are contained in S_r^n, where $k \geq r$. Based on that, the goal of this paper is to establish the following belief.

Claim. A maximal independent connected component of a skeleton represents a primitive *concept* in this collection of documents.

Example 2. *Given a skeleton, S_1^2, of association rules depicted in Figure 2, it is a* 2*-complex composed of the term set V=$\{t_A, t_B, t_C\}$ in a collection of documents. In the skeleton, all* 0*-simplices are neglect, i.e., the terms depicted in dash lines. The simplex set S =$\{\mathrm{Simplex}_1, \mathrm{Simplex}_2, \mathrm{Simplex}_3, \mathrm{Simplex}_4\}$ (*$\mathrm{Simplex}_1$ *is a* 2*-simplex and* $\mathrm{Simplex}_2$, $\mathrm{Simplex}_3$ *as well as* $\mathrm{Simplex}_4$ *are* 1*-simplices) represents generated frequent item-sets from V, and W =$\{w_{A,B}, w_{C,A}, w_{B,C}, w_{A,B,C}\}$ denote their corresponding supports.*

This complex is also a pure 2*-simplex, i.e. triangle, with one maximal independent connected component. The boundary of* 2*-H-simplex has four* 0*-faces (*0*-simplexes) and three* 1*-faces (*1*-simplexes). Since all the simplexes are in the complex, it is a* closed *complex. Therefore, we can say this complex represent a concrete concept. In general, the n-simplex has the following geometric property.*

Property 1. The boundary of a n-H-simplex has $n+1$ 0-faces (vertices), $\frac{n(n+1)}{2}$ 1-faces (edges), and $\binom{n+1}{i+1}$ i-faces ($i \leq n$), where $\binom{n}{k}$ is a binomial coefficient.

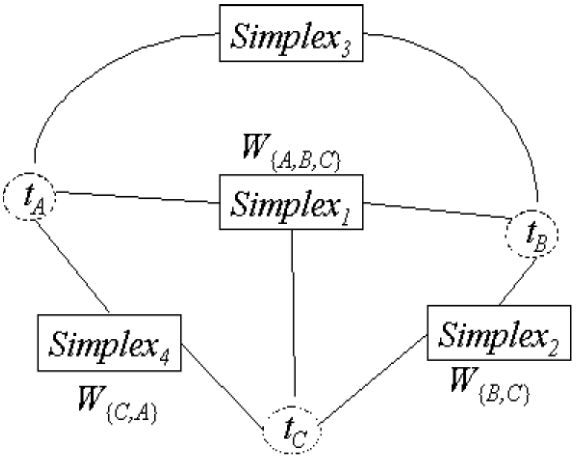

Fig. 2. A simple skeleton S_1^3 of example is composed of three terms $\{t_A, t_B, t_C\}$ from a collection of documents, where each simplex is identified by its tfidf value and all 0-simplices have been removed (the nodes are drawn by using dash circles)

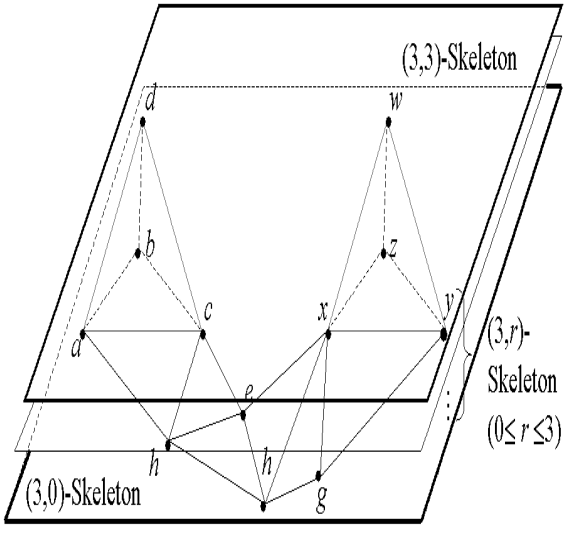

Fig. 3. A simple latent semantic network with its hierarchical structures is generated from Example 1. Obviously the skeleton $(3, 3)$-Skeleton at the topmost layer composed of two maximal connected components as two distinct concepts $\mathrm{Simplex(a, b, c, d)}$ and $\mathrm{Simplex(w, x, y, z)}$ is contained in the skeleton at the lower layer. Except the topmost layer, all the concepts are in some sort of vague discrimination. The bottom layer contains only one connected component, which is a 3-complex. All the concepts are mixed together that make several primitive concepts are non-distinguishable in this connected component.

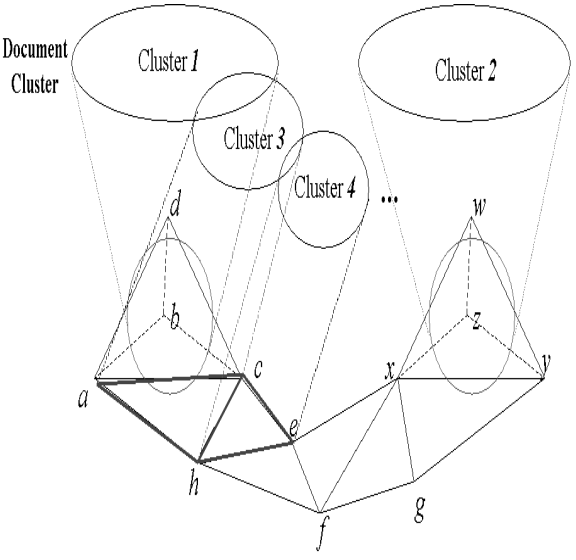

Fig. 4. Each cluster of documents is identified by a maximal connected component. Some cluster may overlap with other cluster because of the common face between them.

This geometric representation properly satisfies the *apriori* property of association rules: if the support of an item set $\{t_1, t_2, \cdots, t_n\}$ is bigger that a minimum support, so are all the non-empty subsets of it. In a complex, the universe of vertices organizes 1-simplices, i.e., frequent 1-itemsets, the universe of 1-simplex represents all possible frequent 1-itemsets and frequent 2-itemsets, and so on.

According to Example 1, it is obvious that simplices within the higher level skeleton S_r^n is contained in the lower level skeleton S_k^n with the same n-complex, $r \geq k$. Figure 3 shows the network hierarchy of the example, each skeleton is represented as a layer. For the purpose of simplicity, skeletons induced from r-complex, in which $0 \leq r < 3$, are neglected. The most distinct concepts of all (without a common concept between them) are existed in the topmost layer, although they could be empty concepts, which means there does not exist any non-overlapped concepts. In this example, the H-simplices $\mathrm{Simplex(a, b, c, d)}$ and $\mathrm{Simplex(w, x, y, z)}$ are two *maximal independent connected components* that demonstrate two discriminating primitive concepts. The H-simplices at the lower layers could have a common face between them. Therefore, the concepts denoted by those H-simplices are vague discriminated as shown in Figure 4 in that an overlapped concept induced by a common face is existed. As seen in the skeleton S_1^3, the maximal connected components generated from simplex $\mathrm{Simplex(a, b, c, d)}$ and simplex $\mathrm{Simplex(a, c, h)}$ have a common face $\mathrm{Simplex(a, c)}$ that makes some documents not able to properly discriminated in accordance with the generated association rules from term a and term c, so are the other maximal connected components in the skeleton. Because of the intersection produced by such subsimplices, some documents would be vague classified into two clusters. The lower the skeleton layer is, the serious the concept overlapping situation is.

4 Finding Maximal Connected Components

For the context of latent semantic ideas within a collection of documents, it is naturally that some similar concepts would be cross-referenced among the collection, especially for a collection of homogeneous documents. Therefore, some professional used words or phrases are often taken to denote a specific idea. No doubt that we can identify them by the usage of those terms. As we already known the best way to recognize them is according to term-term inter-relationships, which are term associations. Following the above statement, combinatorial geometry based latent semantic networks are the perfect model for illustrating the concepts in a huge variety of high-dimensional data, such a document collection. The algorithm for finding all concepts, i.e., maximal connected components, which is generated from the co-occurred terms in a collection of documents, will be introduced as follows.

4.1 Data Structure

In order for the further discussion on the algorithm, let us make the following definitions of the use of geometric notations to represent latent semantic networks on association rules.

Definition 9. *In a latent semanic network, let $\mathcal{V}$ be the set of single terms in a collection of documents, i.e., 0-simplices, and $\mathcal{E}$ be the set of all r-simplices, where $r \geq 0$. If $\mathrm{Simplex}_{\mathrm{A}}$ is in $\mathcal{E}$, its support is defined as $w_($Simplex$_{\mathrm{A}})$, i.e., the* tfidf *of all terms in $\mathrm{Simplex}_{\mathrm{A}}$ co-occurred in a collection of documents.*

A network, which is a complex in geometry, can be represented as a matrix.

Example 3. *As seen in Example 2, the 2-simplex of the network is the set $\{t_A, t_B, t_C\}$, which is also the maximal connected component that represents a primitive concept in a document collection. As Venn diagram, the incident matrix I and the weighted incident matrix I_W of the network are as follows.*

$$I = \begin{pmatrix} 1\,0\,1\,1 \\ 1\,1\,1\,0 \\ 1\,1\,0\,1 \end{pmatrix}.$$

$$I_W = \begin{pmatrix} w_{A,B,C} & 0 & w_{A,B} & w_{C,A} \\ w_{A,B,C} & w_{B,C} & w_{A,B} & 0 \\ w_{A,B,C} & w_{B,C} & 0 & w_{C,A} \end{pmatrix}.$$

The rows correspond to the terms and the columns correspond to the simplices.

Each simplex denotes a connected component, i.e., an undirected association rules. If the simplex is a maximal connected component, it defines a maximal frequent itemset. The number of terms in this connected component defines its *rank*, that is, if its rank is r it is equivalent to frequent $r + 1$-itemsets.

4.2 Algorithm

As we already known, a r-H-simplex denotes a r-connected component, which is a frequent $r+1$-itemset. If we say a frequent itemset I_i identified by an H-simplex $\mathrm{Simplex_i}$ is a subset of a frequent itemset I_j identified by $\mathrm{Simplex_j}$, it means that $\mathrm{Simplex_i} \subset \mathrm{Simplex_j}$. An H-simplex $\mathrm{Simplex_i}$ is said to be a maximal connected component if no other H-simplex $\mathrm{Simplex_j} \in \mathcal{E}$ is the superset of $\mathrm{Simplex_i}$ for $i \neq j$. Documents can be automatically clustered based on all maximal connected components. It provide a soft-computing that allows overlapped concepts exist within a collection of documents.

All connected components are convex hulls, the intersection of connected components is nothing or a connected component. It would induce an vague region for concept discrimination if the intersection is a non-empty simplex. This common face will induce an unspecified concept be-tween them as we have mentioned before. It is not necessary to consider this common face because it has been considered in its super-simplices.

Example 4. *As shown in Figure 5, one component is organized by the H-simplex* $\mathrm{Simplex_1}$=$\{\mathrm{t_A}, \mathrm{t_B}, \mathrm{t_C}\}$*, the other is generated by the H-simplex* $\mathrm{Simplex_5}$=$\{\mathrm{t_C}, \mathrm{t_D}, \mathrm{t_E}\}$.

The boundary of a concept defines all possible term associations in a document collection. Both of them share a common concept that can be taken as a 0*-simplex* $\{t_C\}$*, which is an* 1*-item frequent itemset* $\{t_C\}$.

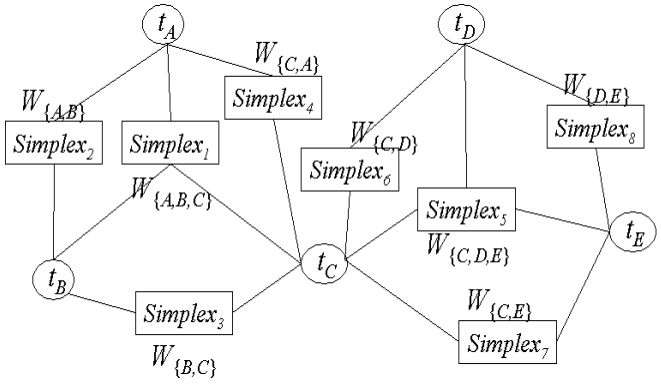

Fig. 5. A complex is composed of two maximal connected components generated by two 2-H-simplices $\mathrm{Simplex(t_A, t_B, t_C)}$ and $\mathrm{Simplex(t_C, t_D, t_E)}$. Both of them contain a common face $\mathrm{Simplex(t_C)}$ that produces an undiscriminating concept region.

Property 2. The intersection of concepts is nothing or a concept that is a maximal H-simplex belonging to all intersected concepts.

Since there is at most one maximal H-simplex in the intersection of more than one connected components and the dimension or rank of the intersection is lower than all intersected simplices. It is convenient for us to design an efficient algorithm for

documents clustering based on all maximal connected components in a complex skeleton by skeleton. It does not need to traverse all complex.

5 Demonstration 1 - Graphical Display of Functions of the NOD2 Genes

Demonstration were performed on the relevant *GeneRIF (References into Function)* document set, publicly available in the EUtils web service of the NCBI Entrez site. Our geometrical model is suitable for representation of biological information, where hierarchical concepts in different complexity could be explored interactively according to the context of application and the various needs of the researchers.

The biological background of the experiment is briefly described here, with the terms or the concepts quoted. “CARD15” gene was found equivalent with “NOD2” gene in recent years. This CARD15/NOD2 gene was discovered associated with inflammatory bowel diseases (“IBDs”) in 2000, and vigorous correlation studies were performed to elucidate the position on the genome or several candidate “chromosomes”. The pathogenesis was proposed later to be “barrier” break in the intestinal (“mucosa”) defense

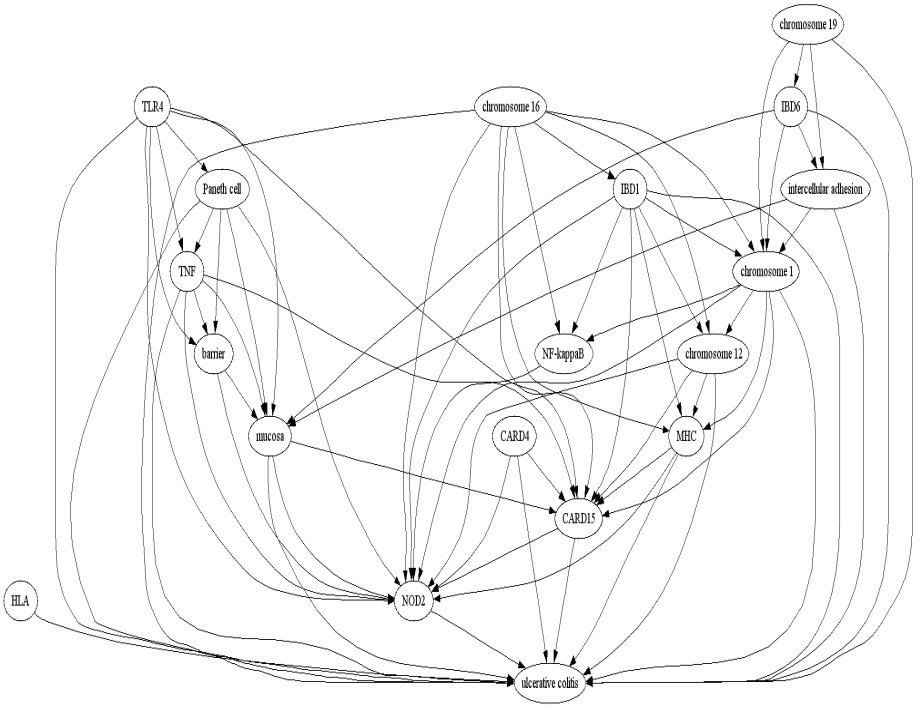

Fig. 6. Functional profiles of the CARD15 gene, rendered by Graph-Viz. The direction of edges are based on TFIDF weighting in this implementation. Our model does not imply directed association.

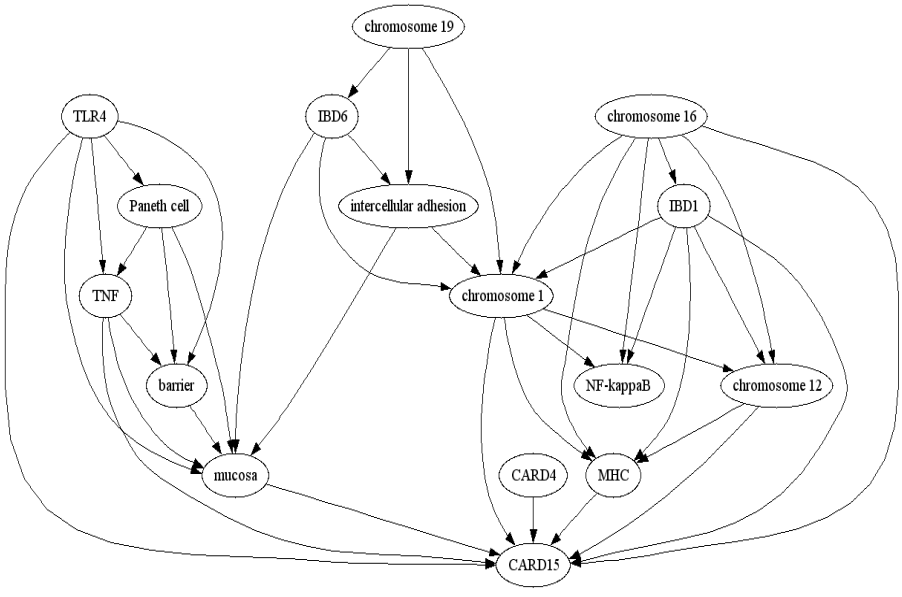

Fig. 7. Functional profiles of the CARD15 gene, with the threshold of the co-occurrence between concept raised. Three biologically meaningful clusters formed.

mechanism due to the genetic defect, then the focus of researchers shifted to the functioning domain of "inflammation" – "TNF", "TLR4", "NF-KappaB", and "Paneth cell".

The GIF document set of CARD15 gene was queried. The abstracts were retrieved, and the important keywords and synonyms were processed by a dictionary derived from UMLS thethaurus. The co-occurrences between the terms were calculated, weighted by TFIDF measurements. In this implementation, the term nodes were ranked by TFIDF weighting, and directed graphes were displayed for additional arrangement of the terms after suggestion by medical domain experts. Our model does not imply directed association.

The nodes of relevant concepts were rendered by the default setting of ATT GraphViz, the layout algorithm of which was according to geometrically even distribution of the nodes and their edges. The nodes with more interconnections or edges were positioneded together, compatible with the clusters of concepts in our model.

In Figure 6, the whole picture of term co-occurrence was shown. In Figure 7, the threshold of visible co-occurrence (the support) was raised, to show the 4-H-simplex or 5-H-simplex concept clusters. Three groups of 4-connected components or 5-connected components were shown in the left, the midlle, and the right regions, corresponding to the concept clusters of the new focus of "inflammatory process" and the older topics and genetic association and chromo-some localization.

The left "inflammatory process" cluster was the 5-frequent itemset with "TLR", "Paneth cell", "TNF", "barrier", and "mucosa". The middle and right clusters were two 4-H-simplex, connected by the intersection of the "chromosome 1" node.

A general concept is not good for classifying/clustering documents. A specific concept can achieve a good precision for document clustering, however, according to coordinate terms, some documents are unable to cluster into a same category by using different terms.

6 Demonstration 2 - Online "Clustering Search" of the NOD2 Gene

An online "clustering search" engine has been constructed based the methodology, which is available at http://ginni.bme.ntu.edu.tw. For use both on biomedical applications and on general purpose, our search engine receives query results (snippets) from PubMed or from Google. We described the PubMed "clustering search" in details, and those readers interested could verify our methodology by various general online searches. For feature extraction, a natural language processing package, QTAG, based on probabilistic parts-of speech natural language processing package was used, the phrase extraction were further fine-tuned by phrase patterns, stop words in some domains, and the weighing of the appearance in the title location. The extracted phrases were selected based on TFIDF measures. For example, the query of "NOD2" gene in PubMed returned 4831 phrases in 200 articles. After the selection by TF*IDF threshold of 0.1, 558 phrases remained to be clustered.

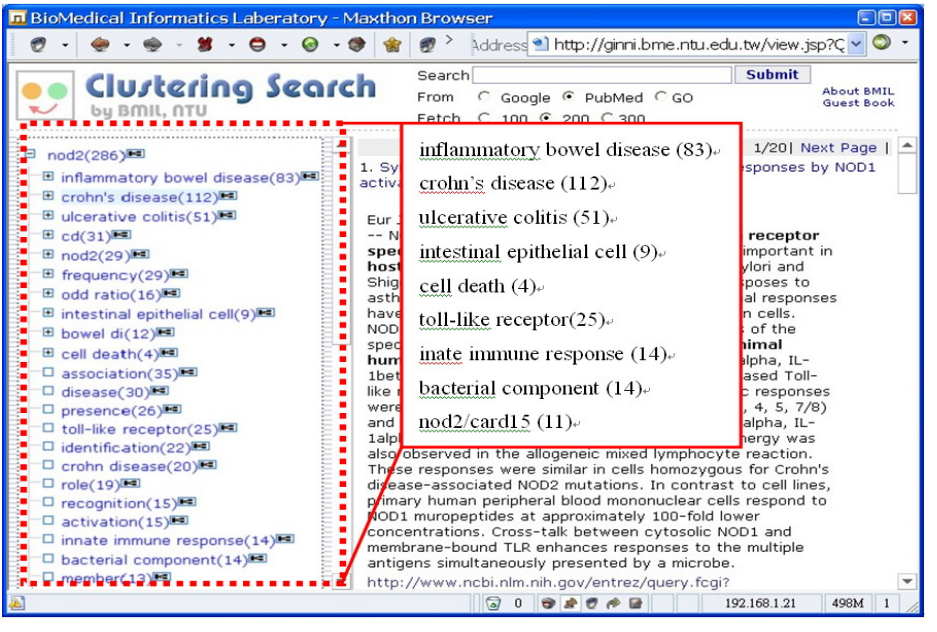

Fig. 8. The initial text browser that shows the hierarchical clustering results and the first nine clusters

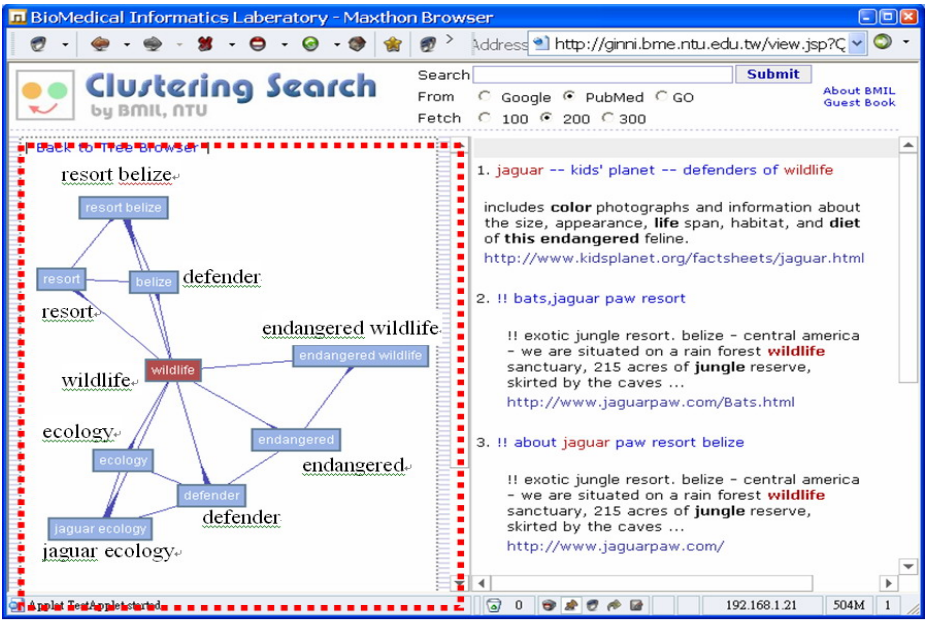

Fig. 9. Graph browser that shows the relations of features in the subgroup "leg ulcer"

Hierarchical clustering was achieved by association rules, which use support and confidence as similarity measure. Support denotes the ratio of the targets in the document corpus, and implies the significance of the association. Confidence is the support of antecedent and consequent divided by the support of the antecedent of the rule, and represents the comparative significance between two phrases. The assumption is that the documents of the same topic are expected to share more common itemsets, the simplicial complices in our model.

The algorithm is:

1. Define the initial value of support and confidence;
2. Find all association rules of pair features;
3. Choose the feature which is pointed by most other features, as the root of the subgroup;
4. Choose other features pointing to the root in this subgroup;
5. Recursively go to Step 2 until the number of the features in the subgroup is lower than a threshold.
6. Choose the feature pointed by most other features from the rest features in the pool, and go to Step 2 until there is no feature in the pool.

The result of the PubMed literature of the NOD2 gene was presented. Figure 8 was the initial screen showing the hierarchical results of clustering in text tree form.

In Figure 9, the graphical view of the details of network relationships between significant concepts in each cluster, user could drag the concept nodes of interest to change the focus and the arrangement to drill on the knowledge structure.

7 Conclusion

Polysemy, *phrases* and *term dependency* are the limitations of web search technology [5]. In the biomedical queries and concept analysis, the problem becomes more severe.

A group of solid term associations can clearly identify a concept. Most methods no matter what is *k-means*, *HCA*, *AutoClass* or *PDDP* classify or cluster documents from the represented matrix of a set of documents. It is inefficient and complicated to discover all term associations from such a high-dimensional and sparse matrix. Given a collection of documents, the associations among frequently co-occurring terms in any of the documents define naturally a geometric complex, which can then be decomposed into connected components at various levels and connected components can properly identify concepts in a collection of documents.

The paper presents a noval approach based on finding maximal connected components for clustering of the functional profile of genes. The r-simplexs, i.e., connected components, can represent the concepts in a collection of relevant documents. It illustrates that geometric complexes are a perfect model to denote association rules in text and is very useful for automatic document clustering and concept grouping, as demonstrated in our experiment in the functional analysis of gene-related documents.

References

1. GO Consortium. Go tools: Editors, browsers, general go tools and other tools. http://www.geneontology.org/doc/GO.tools.html, 2004.
2. C. Friedman, P. Kra, H Yu, M. Krauthammer, and A. Rzhetsky. Genies: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(suppl 1):S74–82, 2001.
3. M. Girvan and M. Newman. Community structure in social and biological networks. In *Proceedings of the National Academy of Sciences*, volume 99, page 8271V76, 2002.
4. T. K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1):21V28, 2001.
5. A. Joshi and Z. Jiang. Retriever: Improving web search engine results using clustering. In A. Gangopadhyay, editor, *Managing Business with Electronic Commerce: Issues and Trends*, chapter 4. World Scientific, 2001.
6. P. Kankar, S. Adak, A. Sarkar, K. Murali, and G Sharma. Medmesh summarizer: Text mining for gene clusters. In *Proceedings of the Second SIAM International Conference on Data Mining*, Apr 2002.
7. J. R. Munkres. *Elements Of Algebraic Topology*. Addison Wesley, Reading MA, 1984.
8. H. Shatkay, S. Edwards, W. J. Wilbur, and M. Boguski. Genes, themes and microarrays: Using information retrieval for large-scale gene analysis. In *Proc Int Conf Intell Syst Mol Biol*, volume 8, pages 317–28, 2000.
9. J. D. Wren and H. R. Garner. Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. *Bioinformatics*, 20(2):191–8, 2004.