

Knowledge Based Search Engine: Granular Computing on Web

Tsau Young (T. Y.) Lin and Jean-David Hsu
Department of Computer Science
San Jose State University
San Jose, California 95192, USA
tylin@cs.sjsu.edu

Abstract

Granular Computing is an emerging technology. This paper summarizes its applications to web. A high frequent co-occurring set of keywords is called a keywordset; it represents some concept in the given document set. These concepts forms a simplicial complex of concepts, which is regarded as a knowledge base for a new search engine. The output will be organized into meaningful clusters. A fundamental difference of this approach from semantic web is that the search engine carries the semantic of a website, instead of the website has to be organized in the sense of semantic web, namely, the semantics is machine readable.

1. Introduction

The Web has fundamentally changed the nature of human life in modern society. Its accelerated growth constantly prompts for new ideas and technologies. A mere retrieval of relevant web pages no longer adequately meets our needs. For example, a internet search on "T. Y. Lin" returns half to several million items. With such an amount, it is not feasible to isolate/identify the desirable item or items manually. Google's ranking is a good idea, but its ranking does not necessary fit all purposes. New fresh look at the problems are needed. Though "semantic web" and "web intelligence" have been the popular proposals, We are proposing a granular computing approach that captures some semantic through knowledge bases.

Granular Computing is an emerging technology. This paper summarizes its applications on web. A high frequent co-occurring set of keywords is called a keywordset; it represents some concept in the given document set. These concepts forms a simplicial complex of concepts, which is regarded as a knowledge base for a new search engine. The output will be organized into meaningful clusters. A fun-

damental difference of this approach from semantic web is that the search engine carries the semantic of a website, instead of the website has to be organized in the sense of semantic web, namely, the semantics is machine readable.

2 Granular Computing - Emerging Field

How is the term Granular Computing coined? In the fall 1996, Professor Lotfi Zadeh suggested granular Mathematics(GrM) to be my research area during my sabbatical leave at Berkeley. To limit the scope, the term granular computing (GrC) was used to label the area of research [18]. So GrC is the computable part of GrM. So, at the beginning, roughly GrC is the computable part of granular mathematics.

Traditionally, "How to solve it" [13] Is not be any part of formal mathematics, however, "how to compute" is an integral part of computing. So GrC has included the mathematical/computational problem solving practices.

We will present the concept in two forms in next two subsections.

2.1 Defining GrC by Examples

The formal definition of GrC has been a shifting paradigm. However, the concept has not been "defined" intuitively by a set of examples quite consistently. Here is a list of representative examples that are taken from [12].

1. Many daily things have been routinely granulated into "sub"things; human body is granulated into head, neck, and etc. The notion is intrinsically fuzzy, vague and imprecise. There is no flawless formal model yet; however, we had proposed one in the keynote of GrC2008. That seems mathematically well defined notion. We had model it in Ninth GrC Model using the category of qualitative fuzzy sets(sofsets) [2].

2. The space and time has been granulated intuitively into infinitesimal granules. The idea was known to Zeno (490BC, implicitly in his paradox), Archimedes (287-212BC) and etc. It led to the invention of calculus, topology, and non-standard analysis; we have modeled it in the First GrC Model.
3. The simplest kind of granulation is partition (see glossary). Its algebraic correspondence, equivalence relation, has played an important role in Euclidean Geometry.(300 BC)
4. Heisenberg uncertainty principle states that, in general, neither the momentum nor the position of a particle can be determined simultaneously with arbitrary great precision. In other words, a precise measurement of the momentum can only determine a "neighborhood"(granule) of positions and vice versa. Note that this "determination" defines a binary relation and vice versa. The idea is abstract into Third GrC Model (A GrC model based on a given binary relation)
5. A committee in a human society (a set of human beings) is a granule. Observe that each member may play different roles. By viewing the collection of roles as a relational schema, a committee is a tuple, not necessarily a subset. The idea is modeled in Fifth GrC Model (A GrC model based on a given set of n -nary relations). In the case all relations are symmetric, the Model is reduced Second GrC Model (A GrC model based on a given set of subsets).
6. Zadehs informal definition [17]: "information granulation involves partitioning a class of objects(points) into granules, with a granule being a clump of objects (points) which are drawn together by indistinguishability, similarity or functionality."

2.2 "The" Formal Model for GrC

We had taken an incremental approach to GrC. So GrC has been a shifting formal notion. In the GrC 2008 keynote, this author finally proposed the category based formal model (Eighth GrC Model) to be *The* Formal Model of GrC, since he can use that model to realize all classical examples given above.

The category theory based model (Eighth GrC Model) actually had been formed much early. However, we had not been able to define flawlessly the Concept of Qualitative Fuzzy Sets(Sofsets) [2]. In this August 2008, we finally broke the ice and had a formal definition of Qualitative Fuzzy Sets(Sofsets) that can realize the fuzzy example, granulation of human body. Now, we are in the state that we can realize all the classical examples given in Section 2.1,

by specifying The Eight GrC Model into various specific categories. Let us first recall:

Definition 1 *A category consists of*

1. *A class of objects, and*
2. *A set $Mor(X, Y)$ of morphisms for every ordered pair of objects X and Y , which satisfies certain properties. For this paper, the formal details are not important; we only need the language loosely.*

Example 1 *Here are some examples.*

1. *The Category of Sets: The objects are classical sets. The morphisms are the maps.*
2. *The Category of Sets with binary relations as morphisms: The objects are classical sets. The morphisms are binary relations. This is the Category of Entity Relationships Models.*
3. *The Category of Power Sets: The object U_X is the power set $P(X)$ of a classical set X . Let U_Y be another object, where Y is another classical set. The morphisms are the maps, $P(f) : U_X \rightarrow U_Y$ that are induced by maps $f : X \rightarrow Y$.*

Definition 2 *Let CAT be a given category.*

Category Theory Based GrC Model:

1. $\mathcal{C} = \{C_j^h, h, j, = 1, 2, \dots\}$ *is a family of objects in the Category CAT .*
2. *There are families (which are bags; see glossary) of Cartesian products, $C_1^j \times C_2^j \times \dots$ of objects, $j = 1, 2, \dots$ of various lengths. They are called product objects.*
3. *An n -ary relation object R^j is a sub-object of the product object $C_1^j \times C_2^j \times \dots \times C_n^j$.*
4. $\beta = \{R^1, R^2, \dots\}$ *be a family of n -ary relations (n could vary).*

The pair (\mathcal{C}, β) , called Categorical GrC Model (Eighth GrC model), is the formal model of granulation. Each tuple is a granule.

3 Knowledge Structure in GrC

A GrC Model has four structures: granular, quotient, knowledge and linguistic structures. Though our main concern is the knowledge structure, We will introduce the four structures.

1. **Granular Structure (GrS):** It is the collection of all granules. In the case of partition, GrS is the collection of the equivalence classes.

2. Quotient Structure (QS): If each granule is abstracted into a point and the intersections of granules are abstract to the interactions of points, then such a collection of points is called quotient structure. In the case of partition, the quotient structure is a classical set, called quotient set.

Informally, GrS is a collection of white boxes (the content of granules are visible), while the quotient set is a collection of black boxes (the contents of granules are hidden). The process of abstracting granular structure into quotient structure is called information hiding; see Examples in Section ??.

3. Knowledge structure: By giving each granule (point) in the quotient structure a meaningful symbol then the named quotient structure is called knowledge structure. The knowledge structure provides an intuitive view of the quotient structure; the symbols and interaction among symbols are in sync with the granules (points) and interactions among granules(points).

In the case of n partitions (equivalence relations), the knowledge structure can be arranged into a n -column relational table. In the case of n binary relations, the table has been called binary information table, granular table or topological table [4], [5].

4. Linguistic structure: By giving each granule in the granular structure a word that reflects its meaning. The interactions among these words are reflected implicitly in precisiated natural language. (in knowledge structure, the interactions among symbols are explicitly reflected from the quotient structure) The linguistic structure is the domain of computing with words

4 Simplicial Complex - A Nice Model

By specifying the category to be the category of sets, we have Relational GrC Model (Fifth GrC model). In this model, β is a collection of n -ary relations. Each tuple is a granule By assuming the symmetry for all n -ary relations, we have the Global GrC Model (Second GrC Model). In this model, granules are subsets (commutative tuples are subsets). So β is a partial covering. Further , if we assume β satisfies further constraints: Any subset of a granule is a granule. Then the Second GrC Model is called a simplicial complex.

We will give this model a geometric interpretation. A n -dimensional Euclidean space is a space in which elements can be addressed using the Cartesian product of n sets of real numbers. A unit point is a point which coordinates are all 0 but a single 1, $(0, \dots, 0, 1, 0, \dots, 0)$. These unit points will be regarded as vertices. We will use them to illustrate the notion of n -simplex.

Let us examine the n -simplexes, when $n = 0, 1, 2, 3$. A 0-simplex $\Delta(v_0)$ consists of a vertex v_0 , which is a point in the Euclidean space. A 1-simplex $\Delta(v_0, v_1)$ consists of two points $\{v_0, v_1\}$. These two points can be interpreted as an open segment (v_0, v_1) in Euclidean space; note that it does not include its end points. A 2-simplex $\Delta(v_0, v_1, v_2)$ consists of three points $\{v_0, v_1, v_2\}$. These three points can be interpreted as an open triangle with vertices v_0, v_1 , and v_2 , which does not include its edges and vertices. A 3-simplex $\Delta(v_0, v_1, v_2, v_3)$ consists of four points $\{v_0, v_1, v_2, v_3\}$ that can be interpreted as an open tetrahedron. Again, it does not includes any of its boundaries.

In general, vertices do not have to be points in Euclidean space, they can be any kind of objects. Formally,

Definition 3 A n -simplex, denoted by $\Delta(v_0, \dots, v_n)$, is a set of independent abstract vertices $\{v_0, \dots, v_n\}$. A q -subset of a n -simplex is called a q -face; it is a q -simplex $\Delta(v_{j_0}, \dots, v_{j_q})$ whose vertices are a subset of $\{v_0, \dots, v_n\}$ with cardinality $q + 1$.

More generally, we have ([16][p.108])

Definition 4 A simplicial complex C consists of a set $\{v\}$ of vertices and a set $\{s\}$ of finite nonempty subsets of $\{v\}$ called simplexes such that

- Any set consisting of one vertex is a simplex.
- Closed condition: Any nonempty subset of a simplex is a simplex.

Any simplex s containing exactly $q + 1$ vertices is called a q -simplex. We also say that the dimension of s is q and write $\dim s = q$. We will refer to C as a non-closed simplicial complex, if the closed condition is not fulfilled for all its constituting simplexes.

Any set of $n + 1$ objects can be viewed as a set of abstract vertices. A simplex is said to be maximal if it is not a face of any other simplex. Moreover, if the maximal dimension of the constituting simplexes is n then the complex is called n -complex.

Example 2 In Figure 1, we have a simplicial complex that consists of twelve vertices that are organized in the form of a 3-complex, denoted by S^3 .

Let us enumerate maximal simplexes in S^3 :

1. The maximal 3-simplex $\Delta(a, b, c, d)$, and all its faces.
2. The maximal 3-simplex $\Delta(w, x, y, z)$ and all its faces.
3. The maximal 2-simplexes lying "between" two 3-simplexes: $\Delta(a, c, h)$, $\Delta(c, h, e)$, $\Delta(e, h, f)$, $\Delta(e, f, x)$, $\Delta(f, g, x)$, $\Delta(g, x, y)$ and their faces.

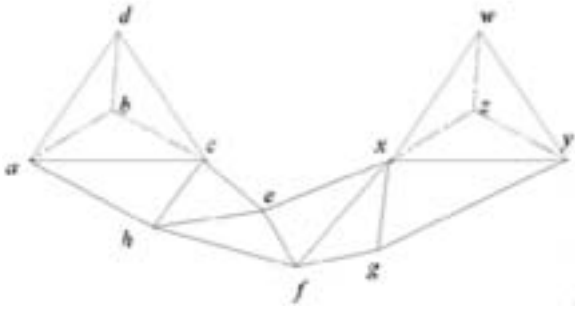


Figure 1. A complex with twelve vertices.

5 Knowledge Complex - Keyword Simplicial Complex

In this section, we will examine a knowledge structure for a simplicial complex that is derived from a document set.

5.1 Term order, Sequence vs Set of Terms

The mathematical model applies well to linear text, vertices being terms and edges being the connection between terms nearby (i.e. the distance can be the number of terms in between 2 terms). Nevertheless, in order to make proper use of this model, we need to introduce some minor adjustments.

In many situations, may need to ignore low dimension simplexes in the Simplicial Complex of linear texts. The idea is to iterate through the Simplicial Complex much like through a hypergraph. Edges with low weight ought to be ignored to help identify "clear cuts" between clusters.

The simplex model has no notion of *vertex order*. The underlying observation is that text comes in a unique order, rarely the other order make sense (if they do, the number of occurrence is very small and will not effect our computations), so we will, with negligible lose, assume the un-ordered set really represents this particular order. Also in some occasions, a set of keywords may represent more than one meanings. We will deal with it at the time of interpreting the keywords.

Next, we consider the impact in truncating sentences. For instance, removing coordinate conjunctions from a text also removes the relationship between the clauses being joined, altering the overall semantic of the text. More obvious one is removing a negation. Our observation is that such kinds of loses are at micro level, does not effect the overall "summarization" of a documents. By these observation, we believe simplex model is a reasonable model for clustering documents.

5.2 Association(rule)s and Keywords

A formal definition of association rule can be found in [1]. For such an association(rule), there are two measurements, SUPPORT and CONFIDENCE. In this paper, we only use SUPPORT value. The SUPPORT of a set of a nearby keywords is the percentage of documents in the document set that contain the keywords. An association (or frequent itemset) is a set of keywords $K = \{k_1, k_2, \dots, k_n\}$ in a set of document $D = \{d_1, d_2, \dots, d_i\}$, where $d_i = \{k_{i1}, k_{i2}, \dots, k_{ij}\}$ and $k_{ij} \in K$, whose SUPPORT is greater than or equal to the threshold value. We will use associations as we are more interested in the semantic of the interactions among the keywordset.

Recall an important notion from data mining, the key property of Agrawal's associations [1] is the Apriori condition.

It can be stated as follows.

Definition 5 *Apriori condition: Any q -subset of a n -keywordset is a q -keywordset for $q \leq n$.*

where a q -keywordset is a keywordset composed of q keywords.

The search of a n -keywordset does not start until its Apriori condition has been met.

Similarly, the simplicial complex has a property called *closed condition* (cf Definition 4) that states that any subset of a n -simplex is itself a simplex in the simplicial complex. **A keyword can be regarded as an abstract vertex and consequently a $(q + 1)$ -keywordset can be interpreted as an abstract q -simplex of keywords.**

5.3 Keyword Simplicial Complex

The following lemma is immediate.

Lemma 1 *The closed condition of abstract simplexes is the Apriori condition of keywords and vice versa.*

With this lemma, we have the following theorem:

Theorem 1 *The pair (V_{text}, S_{text}) is an Abstract Simplicial Complex, where*

1. V_{text} is the set of keywords, and is regarded as a set of abstract vertices, called keyword-vertices.
2. S_{text} is the set of keywordsets (associations) and is regarded as a set of abstract simplexes, called keyword-simplexes.

We may call this simplicial complex the Keyword Simplicial Complex (KSC).

Definition 6 *The keywords and keywordsets in V_{text}, S_{text} are names of vertices and simplexes respectively. In other words, each granule is named, So KSC is the knowledge structure of the simplicial complex model.*

Theorem 2 *Let A and B be two document sets written into different languages. The simplicial complexes of A and B are isomorphic.*

This theorem simply tells us that using the simplex analogy, we can determine if two sets are the translation of one another.

6 A Knowledge Base on Web

A document or a web page can be viewed as a list of linearly ordered **tokens** or **terms**; to emphasize this view, we call it linear text. It is a knowledge representation of human idea or human thoughts. However, mathematically speaking, this is not a good representation, because The semantics in the linear text can only be revealed through human reading.

In this section. We have constructed a mathematical structure that reveal authors idea or thoughts mathematically. We may schematically summarize the flow of constructions; it is a variation of [9]

$$\begin{bmatrix} \text{author's} \\ \text{IDEA} \end{bmatrix} \longrightarrow \begin{bmatrix} \text{linear} \\ \text{text} \end{bmatrix}$$

↓↓(summarize)

$$\text{IDEA} \equiv \begin{bmatrix} \text{human} \\ \text{thoughts} \end{bmatrix} \longrightarrow \begin{bmatrix} \text{keyword} \\ \text{simplicialcomplex} \end{bmatrix}$$

Note that none of those \longrightarrow is a mathematical map; it merely indicates information flows. A novel point in this combinatorial structure is the capability to reveal the idea hidden between lines, namely, the model can read "between the lines."

6.1 Semantics - "Reading between the lines"

Each document describes some ideas in human mind, which may consist of many levels and wide ranges of concepts. We capture those concepts through high-frequency tokens and set of tokens. We will use keywords and keywordsets to denote high-frequency tokens (terms) and high-frequency tokensets respectively, where a high-frequency

tokenset is high-frequency tokens that co-occur within close distance in a document. Following the terminology in association mining, we may also use an association to refer to a keywordset; they will be used interchangeably throughout this paper.

First, let us observe some interesting phenomena. A keywordset, semantically, may have nothing to do with its individual keywords. For example,

1. The keywordset "Wall Street" represents a concept that has nothing to do with "Wall" and "Street"
2. The keywordset "White House" represents an object that has very little to do with "White" and "House."
3. The keyword set "Wall, China, Security" represents a concept, Chinese Wall Security Policy. that is beyond the meaning of three key words, "Wall," "China," and "Security."

These examples indicate that the strength of this approach is the ability to capture the notion that is defined implicitly by the keywordset, in plain words, the capability of "reading between the lines." Let us formally define the keywordset, the mechanism that carries the unspoken concept.

Definition 7 *A n_d -keywordset is a set that has a high number of co-occurrences (SUPPORT) of n keywords that are at most d tokens apart. In the case that d and n are understood, we abbreviate it simply as keywordset.*

Note that a keyword has been informally understood as a high frequency token.

7 Applications-Document Clustering

In traditional clustering, we partition a document set into disjoint groups, namely, equivalence classes of documents. However, many documents are inter-related in some concepts and totally unrelated in others. So we propose a concept based clustering where we use the concepts to group documents.

Recall that a keyword is the name of a vertex, and a keywordset is the name of a simplex; they are all high frequent item or itemsets. Observe that a keyword or keywordset represents a concept. So we say a document contain a concept, if it contains the corresponding keywordset/ Two documents are clustered together if both contains the same concept [6].

Table 1, 2, 3, 4 are extracted from [10]. the Reuters data in the UCI database located at <http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html>). We only collected the high dimensional clusters for clarity, but keep in mind that main clusters may have undisplayed

”weak connections”.

- 1) An I-concept (I means intermediate) is the concept correspond to a simplex.
- 2) An P-concept(P means primitive) is the concept correspond to a maximal simplex.
- 3) A C-concept is the concept correspond to a connected component of the KSC.

Two documents are I/P/C-clustered together if both documents have the same I/P/C-concept. Recall that two keywordsets, w_0 and w_n , are in the same connected component if there is a finite sequence w_0, w_1, \dots, w_n of keywordsets, such that any consecutive two keywordsets have a common sub-keywordset.

Each tuple in Table 1, 2, 3, 4 represents a cluster, called it a P-cluster, where P-concepts are used for clustering. Column A enumerates P-clusters. Column C indicates the number of documents in this P-cluster. The remaining columns list the keywords of this P-concept.

Notice that P-clustering may not be desirable in all occasions. For example the first 7 tuples in Table 1 have a substantial number of common faces among the maximal simplexes, so many documents refer to a lot of common concepts: These documents have many common concepts and are therefore hardly differentiable by P-concept.

Based on these definitions, the documents in the TABLES only show one connected component. This clustering is not very satisfactory because it contains groups that are *nearly* disjoint. They are grouped together simply due to some discussion(s) on very general or equivalently very low-level I-concepts (corresponding to 1-simplexes or vertexes).

Therefore, in certain situations, we may decide to ignore some I-concepts. we will call this method relative clustering.

4) Relative Clustering by pair (C-concept, sub-concept)

We will label each document by the given pair if the document contains any I-concept within the C-concept that is not within the sub-concept. Here, sub-concept corresponds to a sub-complex of the connected component. In plain words, two documents are not going to be clustered together if their only common topics are in the sub-concept.

Consider the sub-complex that consists of 1-simplex $\Delta(dr, mn)$, 0-simplex $\Delta(ct)$, and all their faces. The relative clustering can *nearly* be expressed in the new tables, where three columns, ”ct”, ”dr” and ”mn,” must be removed. Note that this is equivalent to removing $\Delta(ct, dr, mn)$ and all its faces. This approach is a bit greedy and therefore, in some instances, we may refer to

this approach as a sloppy interpretation. Nevertheless, in our example the ”cuts” are clear and the resulting table provides four completely disjoint clusters. Each cluster can now be assigned a C-concept.

8 Conclusion

This series of investigations was triggered by Professor I-Jen Chiangs enquiry while I was visiting IIS, Academia Sinica in Taiwan. His query trigger my association of hypergraph to simplicial complex [6]. Though the simplicial complex, by definition, is a subset of hypergraph, yet the fundamental spirit is very different. Hypergraphs are graphs, so the vertices are the vital entities, while in simplicial complex, the main objects are the simplexes (corresponding to hyper edges), vertices are merely low dimensional things that can be discard at any time. However, at the beginning, we investigated both and tries to apply to clustering the document sets as have done by the group in Minnesota.

This approach is very different from the classical approach. The clustering is concept based. The documents are clustered into groups, each group discusses different concepts. Various levels of granularity or details can be considered by excluding various subcomplexes - relative clustering. This paper is merely an overview of the process, the details will be reported somewhere else. The search engine is also concepts based index, so its output is very different from current search engine.

References

- [1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th VLDB Conference*, 1994.
- [2] Lin, T. Y.(1996): A Set Theory for Soft Computing. In: *Proceedings of 1996 IEEE International Conference on Fuzzy Systems*, New Orleans, Louisiana, September 8-11, 1996, 1140-1146.
- [3] Lin, T.Y. (1998a): Granular Computing on binary relations I: data mining and neighborhood systems. In Skoworn, A. and Polkowski, L., editors, *Rough Sets In Knowledge Discovery*, pages 107–121. Physica-Verlag, Heidelberg
- [4] Lin, T.Y. (1998b) Granular Computing on Binary Relations II: Rough set representations and belief functions. In Skoworn, A. and Polkowski, L., editors, *Rough Sets In Knowledge Discovery*, pages 121–140. Physica-Verlag, Heidelberg
- [5] Lin, T. Y. Liao, C. J (2005): Granular Computing and Rough Sets. In: Oded Maimon, Lior Rokach (Eds.):

A	B	C	1	2	3	4	5	6	7	8	9	10	11	12
0	1	29	15	apri	march		pai		record	ct	31			
1	1	65	15	apri	march	prior	pai	reuter	record	ct				
2	1	26	15		march	prior	pai	reuter	record	ct	31			
3	1	23	15	apri		prior	pai	reuter	record	ct			30	
4	1	53		apri	march	prior	pai	reuter	record	ct		10		
5	1	47		apri	march	prior	pai	reuter	record	ct				20
6	1	43		apri	march	prior	pai	reuter	record	ct	31			
7	2	32												
8	2	46												
10	2	34												
11	2	21												
12	2	22												
13	2	34												
14	3	20												
15	3	20												
16	4	42								ct				
9	4	181								ct				

Table 1. A: Document-Group number; B: Relative Connected Component number (the sub-concept is a complex of $\Delta(dr, mn)$, $\Delta(ct)$ and their faces; C: number of documents; the remaining numbers are token numbers. Conclusion: Component 1&4 have 1 common keyword = ct; Component 2&3 have 1 common keyword = dr; Component 2&4 have 1 common 1-simplex = $\Delta(dr, mn)$

A	B	C	13	14	15	16	17	18	19	20	21	22
0	1	29										
1	1	65										
2	1	26										
3	1	23										
4	1	53										
5	1	47										
6	1	43										
7	2	32	mn	dr		cover			issu	statement		
8	2	46	mn	dr	convert		debentur	due			subordin	2012
10	2	34	mn	dr	convert	cover	debentur	due	issu		subordin	
11	2	21	mn	dr	convert	cover	debentur		issu	statement	subordin	
12	2	22	mn	dr	convert	cover			issu	statement	subordin	
13	2	34	mn	dr	convert		debentur	due	issu		subordin	2012
14	3	20		dr								
15	3	20		dr								
16	4	42	mn									
9	4	181	mn	dr								

Table 2. A: Document-Group number; B: Relative Connected Component number (the sub-concept is a complex of $\Delta(dr, mn)$, $\Delta(ct)$ and their faces; C: number of documents; the remaining numbers are token numbers. Conclusion: Component 1&4 have 1 common keyword = ct; Component 2&3 have 1 common keyword = dr; Component 2&4 have 1 common 1-simplex = $\Delta(dr, mn)$

A	B	C	23	24	25	26	27	28	29	30	31	32
0	1	29										
1	1	65										
2	1	26										
3	1	23										
4	1	53										
5	1	47										
6	1	43										
7	2	32										commiss
8	2	46										
10	2	34										
11	2	21										
12	2	22										
13	2	34										
14	3	20	20	custom	arrang	market	biion	secur	repurchas	agreement		
15	3	20	20	custom	arrang	market	biion	secur	repurchas		govern	
16	4	42	42									

Table 3. A: Document-Group number; B: Relative Connected Component number (the sub-concept is a complex of $\Delta(dr, mn)$, $\Delta(ct)$ and their faces; C: number of documents; the remaining numbers are token numbers. Conclusion: Component 1&4 have 1 common keyword = ct; Component 2&3 have 1 common keyword = dr; Component 2&4 have 1 common 1-simplex = $\Delta(dr, mn)$)

A	B	C	33	34	35	36	37	38	39	40	41	42	43	44
0	1	29												
1	1	65												
2	1	26												
3	1	23												
4	1	53												
5	1	47												
6	1	43												
7	2	32							exchang					registr
8	2	46	offer											
10	2	34												
11	2	21												
12	2	22												registr
13	2	34												
14	3	20												
15	3	20												
16	4	42		oss	profit		oper			rev	net		1000	
9	4	181		oss	profit	shr		year			net			

Table 4. A: Document-Group number; B: Relative Connected Component number (the sub-concept is a complex of $\Delta(dr, mn)$, $\Delta(ct)$ and their faces; C: number of documents; remaining numbers are token numbers. Conclusion: Component 1&4 have 1 common keyword = ct; Component 2&3 have 1 common keyword = dr; Component 2&4 have 1 common 1-simplex = $\Delta(dr, mn)$)

The Data Mining and Knowledge Discovery Handbook. Springer 2005, 535-561. ISBN 0-387-24435-2.

- [6] T. Y. Lin and I-Jen Chiang "A simplicial complex, a hypergraph, structure in the latent semantic space of document clustering, International Journal of Approximate Reasoning, 2005
- [7] T. Y. Lin." Granular Computing: Examples, Intuitions and Modeling." In: the Proceedings of 2005 IEEE International Conference on Granular Computing," July 25-27, 2005, Beijing China, 40-44.
- [8] T. Y. Lin." Granular Computing II: Infrastructure for AI-Engineering." In: the Proceedings of 2006 IEEE International Conference on Granular Computing," May 10-12, 2006, Atlanta, Georgia, USA, 2-7.
- [9] Lin T. Y., Sutojo A, Hsu, J-D. (2006): Concept Analysis and Web Clustering using Combinatorial Topology. In: Workshops Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006), 18-22 December 2006, Hong Kong, China. IEEE Computer Society 2006, ISBN 0-7695-2702-7 412-416
- [10] Kevin Lind, SJSU project report, 2006
- [11] <http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html>
- [12] Lin, T. Y.(2008): Granular Computing: Ancient Practices, Modern Theories and Future Directions, The Encyclopedia on Complexity of Systems, to appear.
- [13] G. Polya (1957): How to Solve It, 2nd ed., Princeton University Press, 1957, ISBN 0-691-08097-6.
- [14] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
- [15] G. Salton and C. Buckley. Term Weighting Approaches in Automatic Text Retrieval, 1960, Information Processing and Management, 24, Vol 5, 513-523
- [16] Spanier, E. H (1966): Algebraic Topology, McGraw Hill, 1966 (Paperback, Springer, Dec 6, 1994)
- [17] Zadeh, L. A. (1996): The Key Roles of Information Granulation and Fuzzy logic in Human Reasoning. In: *1996 IEEE International Conference on Fuzzy Systems*, New Orleans, Louisiana, September 8-11, 1, 1996.
- [18] Zadeh, L.A. (1998): Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/ intelligent systems, *Soft Computing*, 2, 23-25.