Concept Mining using Association Rules and Combinatorial Topology

Albert Sutojo
San Jose State University, CA, USA
asutojo@email.sjsu.edu

Abstract— **The collection of concepts in a document set can be represented by a geometric structure called simplicial complex of combinatorial topology where each keyword is represented as a vertex and the relation between keywords as simplex. A simplex which consists of more than one keyword is a high-frequency keywordset. These keywords occur close to each other within a document which also occur frequently within a set of documents. The high frequent occurence of these keywords shows relations between keywords. These relations carry concepts. The relations of these keywords can be captured by Association Rule Mining and represented as simplices. The collection of all these simplices, represents the structure of concepts within a document set. Based on this topology, documents are clustered and the collection of simplices can serve as document index.**

I. Introduction

Data Mining is about finding interesting and useful patterns from data. The data can be structured data like in database tables, or unstructured data like in web documents. This paper proposed a technique to extract concepts within a set of documents. The paper was based on a master thesis by the same author [9].

Terms that appear many times in different documents tend to have particular meaning. A set of more than one keyword that co-occurs in a document within a close distance might carry an association. This association carries a particular meaning as well. This terms association might create a particular concept beyond the meaning of the individual keywords. A set of co-occuring keywords might consists of two keywords , three keywords or more. Besides mining single terms, sets of more than one terms can also be extracted from documents. These sets of terms association can be represented by a geometric structure called simplical complex of combinatorial topology. Documents can be clustered based on this topology.

The approach proposed in this paper uses the Term Frequency Inverse Document Frequency calculation method, which is known as TFIDF, to extract keywords from documents. The approach also uses Association Rule Mining to find terms association which consists of high-frequency terms that co-occur within a close distance in a document and these terms associations are represented as simplicial complex structure.

The structure of this paper is organized as follows : section 2 introduces the supporting theories and basic concepts of TFIDF and association rule mining. Section 3 describes the concept of simplicial complex structure used to represent terms association. Section 4 shows some experimental results. Section 5 presents the conclusion and future enhancement.

II. Supporting Theories

A. TFIDF

The TFIDF or Term Frequency Inverse Document Frequency method measures the significance of a term in a document within a set of documents. The idea of TFIDF is to discriminate terms by occurence within a set of documents. Terms that occur in almost every document tend to be common terms or stop words and do not convey important meaning. Terms that occur only in some documents might be considered as important terms, or keywords, which convey particular meaning. In this paper terms and keywords are use interchangeably.

Many formulas to calculate TFIDF values have been proposed. We choose the following formula to calculate TFIDF values :

Definition 2.1: Let T_r denote the total number of documents in the collection. We approximate the significance of a token t_i in a document d_j, itself in T_r, by its TFIDF value. It is calculated as

$$\mathrm{TFIDF}(\mathrm{t_i}, \mathrm{d_j}) = \mathrm{TF}(\mathrm{t_i}, \mathrm{d_j})\mathrm{IDF}(\mathrm{t_i})$$

where $\mathrm{TF}(\mathrm{t_i}, \mathrm{d_j})$ stands for Term Frequency and is defined by

$$\mathrm{TF}(\mathrm{t_i}, \mathrm{d_j}) = \begin{cases} \frac{N(t_i, d_j)}{|d_j|} & \text{if } N(t_i, d_j) > 0 \\ 0 & \text{otherwise} \end{cases}$$

where t_i is a term of document d_j, $N(t_i, d_j)$ denotes the frequency t_i in d_j, and $|d_j|$ is the total number of tokens in document d_j.
$\mathrm{IDF}(\mathrm{t_i})$ stands for Inverse Document Frequency and is defined as

$$\mathrm{IDF}(\mathrm{t_i}) = \log \frac{\mathrm{T_r}}{\mathrm{df}(\mathrm{t_i})}$$

where $\mathrm{df}(\mathrm{t_i})$ stands for Document Frequency of term t_i and denotes the number of document T_r in which t_i occurs at least once.

Terms that occur in a large number of documents tend to be stop words. By using $\mathrm{TF} \times \mathrm{IDF}$ [6], [7] calculation, it is expected that stop words can be discriminated by their TFIDF score. Based on the characteristics of the logarithmic function of $\frac{T_r}{\mathrm{df}(\mathrm{t_i})}$, stop words tend to have TFIDF scores close to or equal to zero.

0-7695-3032-X/07 $25.00 © 2007 IEEE
DOI 10.1109/GrC.2007.154

Note that the TFIDF value only reflects the importance of a terms in one particular document. Its value is local to each document. It does not measure the overall significance of a token in the set of documents; Moreover, the $\mathrm{IDF(t_i)}$ value is at its highest when the token appears in only one document. Therefore, the opinion here is that selecting keywords solely based on TFIDF values is not enough. Words that have low TFIDF values are not good terms. Neither are terms that occur only in very few documents. This leads to our definition of keywords.

Definition 2.2: Keywords are considered important if they have high TFIDF value *and* high *Document Frequency* (DF) value. We say that DF value is high if it is greater than a given threshold value.

Typically, the TFIDF value is an indicator to identify keywords, and the DF value is an indicator to identify interesting keywords.

B. Co-occuring Keywords

Each document describes some ideas of the human mind, which may consist of many levels and wide ranges of concepts. A set of one keyword describes a concept. However, a set of more than one keyword might describe a new meaning that is beyond the meaning of each individual keywords. If this new meaning is important, the set of these keywords will consistently appear in the form of co-occuring keywords. Concepts within documents can be captured through high-frequency and co-occuring keywords.

The term keywords and keywordset will be used to denote high-frequency keywords and high-frequency keywordset respectively. High-frequency keywordset is high-frequency keywords that co-occur within a close distance in a document.

Interestingly, a keywordset, semantically, may have nothing to do with its individual keywords. For example,

1) The keywordset “Wall Street” represents a concept whose meaning is beyond the word “Wall” and “Street.”
2) The keywordset “White House” represents an object that is different than the words “White” and “House.”

These examples indicate that the strength of this approach is the ability to capture the meaning that is defined implicitly by the relation of terms or keywords .

Keywordset, the mechanism that carries the unspoken concept, can be formally defined as :

Definition 2.3: An n_d-keywordset is a set that has a high number of co-occurrences of n keywords that are within at most d tokens apart. In the case that d and n are understood, we abbreviate it simply as a keywordset.

C. Association Rule for Mining Keywordsets

Association Rule Mining [1] is used to show relationships between keywords. Interesting and important keywords occur frequently enough in a document set. Associations between these keywords create semantics beyond the meaning of the individual keywords. In the context of mining keywordsets association, the association rule can be defined as follows :

Definition 2.4: Given a set of keywords K $=\{k_1, k_2, ... \, k_n\}$ and a set of document D $=\{d_1, d_2, ...d_i\}$ where d$_i$ $=$ $\{k_{i1}, k_{i2}, ...k_{ij}\}$ and $k_{ij} \in K$, an association rule is an implication of the form A $\Rightarrow$ B where A,B are sets of keywords called keywordsets and $A \cap B = \emptyset$.
Interest usually lies in the important association rule. The importance of an association rule is measured by two features called SUPPORT and CONFIDENCE.

Definition 2.5: SUPPORT for an association rule A $\Rightarrow$ B is the percentage of documents in the document set that contain keywordsets $A \cup B$ greater or equal than the threshold value.

Definition 2.6: CONFIDENCE for an association rule A $\Rightarrow$ B is the ratio of the number of documents that contain keywordsets $A \cup B$ to the number of documents that contains A. Simply, it is the ratio between SUPPORT($A \cup B$) and SUPPORT(A).

In detecting interesting keywords within a set of document, we only look at the SUPPORT value. More precisely, interesting keywordsets are high-frequency keywords that have high TFIDF values and high SUPPORT values and they co-occur within a close distance in a document.

Many algorithms have been developed to find association rules. The most popular algorithm is The Apriori Algorithm. The algorithm generates n-keywordsets from $n-1$-keywordsets, where $n > 1$. This means that the process of generating n-keywordsets depends on the previous step when generating $n-1$-keywordsets.

One notable characteristic of the Apriori Algorithm is the **Apriori Condition**, from which the name of the algorithm was taken. The condition states that if n-itemset is frequent, then $n-1$-itemset must also be frequent. This property of the Apriori Algorithm reduces the search space required in generating n-keywordsets. This condition is similar to that of simplicial complex structure in combinatorial topology.

III. Simplicial Complex of Combinatorial Topology

This section defines some technical terms from combinatorial topology where the concept of n-simplex and simplicial complex are introduced.

A. n-Simplex

An 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 whose coordinates are all 0 except for a single 1, $(0, \ldots, 0,\ 1,\ 0, \ldots,\ 0)$. These unit points will be regarded as vertices. They will be used to illustrate the notion of n-simplex.

Let us examine the n-simplices, 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 the 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, that does not include the edges and vertices. A 3-simplex $\Delta(v_0, v_1, v_2, v_3)$ consists of four points $\{v_0,\ v_1,\ v_2,\ v_3\}$ and can be interpreted as an open tetrahedron. Again, it does not include any of its boundaries.

Definition 3.1: A n-simplex, denoted by $\Delta(v_0, \ldots, v_n)$, is a set of independent abstract vertices $\{v_0, \ldots, v_n\}$. A q-subset of a n-simplex is called a q-face; it is a q-simplex $\Delta(v_{j_0}, \ldots, v_{j_q})$ whose vertices are a subset of $\{v_0, \ldots,\ v_n\}$ with cardinality $q + 1$.

B. Simplicial Complex

The set V is called the *vertex set* of the complex *K*. Each *p*-simplex is said to be of dimension *p*. The largest integer *n* for which $\sigma_n\ \epsilon\ K$ is called the *dimension* of *K*.

A simplicial complex can be defined as

Definition 3.2: A simplicial complex C consists of a set $\{v\}$ of vertices and a set $\{s\}$ of finite nonempty subsets of $\{v\}$ called simplices such that

- Any set consisting of one vertex is a simplex.
- Any nonempty subset of a simplex is a simplex (closed condition).

Any simplex s containing exactly $q + 1$ vertices is called a q-simplex. The dimension of s is q and written as dim(s)=q. C will be referred to as a *non-closed simplicial complex*, if the closed condition is not fulfilled for all its constituting simplices.

To re-iterate, 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 simplices is n, then the complex is called n-complex.

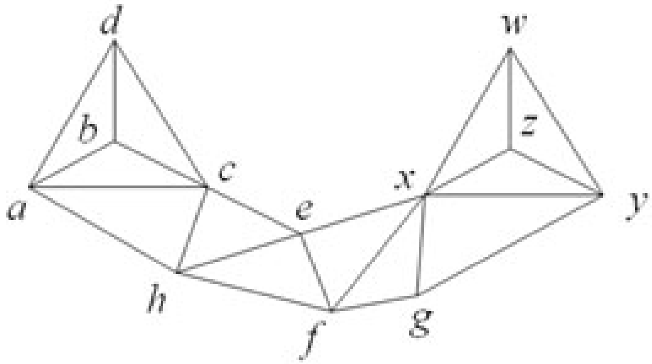


Fig. 1. A complex with twelve vertices.

In Figure 1, a simplicial complex consisting of twelve vertices that are organized in the form of a 3-complex, denoted by S^3.

Maximal simplices in S^3 should be enumerated:

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-simplex lying "between" two 3-simplices: $\Delta(a, c, h)$, $\Delta(c, h, e)$, $\Delta(e, h, f)$, $\Delta(e, f, x)$, $\Delta(f, g, x)$, $\Delta(g, x, y)$ and their faces.

IV. Combinatorial Structure of Linear Text

There is a relationship between the nature of the Apriori Condition and the structure of keywords and keywordsets in simplicial complex. Apriori says if n-keywordsets are frequent, then $n - 1$-keywordsets are also frequent. This can be formally defined as :

Definition 4.1: Apriori Condition : Any q-subset of an n-keywordset is a q-keywordset for $q \leq n$ where a set containing exactly q-elements is abbreviated as a q-subset.

Keywordsets are built from the "bottom-up". n-keywordsets are built from $n - 1$-keywordsets. If n-keywordsets are frequent, then their subset must also be frequent. Surprisingly, there is a condition similar to the Apriori Condition called closed condition that applies to the simplicial complex (Section III). Roughly, it says that any subset of an n-simplex in the simplicial complex 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**. So the following lemma is immediate :

Lemma 4.2: The closed condition of abstract simplices is the Apriori Condition of keywordsets and vice versa.

With this lemma, we have the following theorem :

Theorem 4.3: 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 simplices called keyword-simplices.

This simplicial complex is called a Keyword Simplicial Complex (KSC).

Theorem 4.4: Let A and B be two document sets, where B is a translation of A into another language then the simplicial complexes of A and the simplicial complexes of B are isomorphic.

This theorem is striking: Using this model, we can determine if two sets of documents written in different languages are similar, even without translation.

V. The Meaning of Geometric Structures

This project clusters documents by n-keywordset association, which also known as n-simplex in simplicial complex structure. The followings are some linguistic meaning of this structure. These terms refer to some notions in the document set.

L1) Whole keywords in simplicial complex structure represents the whole **IDEA** of a document set.
L3) Each keyword k represents a basic concept called **B-concept**(k).
L4) A q-simplex Δ represents some intermediate concepts, which is known as **I-concept**(Δ).
L5) A maximal simplex represents a primitive concept called **P-concept**.
L6) A connected component represents a complete concept called **C-concept**.

A link can be established between a document and a P-concept if the document contains the corresponding maximal keyword-simplex. Plainly, two documents are clustered together if both of them have the same maximal keywordset [2]. Maximal keywordset is n-keywordset which is not a subset of $n + 1$-keywordset. These documents have addressed the P-concept.

A link can be established between a document and a C-concept if the document contains a keyword-simplex that is part of the connected component. In other words, C-concept is a collection of P-concepts with all its sub-keywordsets, which are I-concepts. Furthermore, two keywordsets, w_0 and w_n, are in the same connected component if there is a finite sequence $(w_0, w_1, \ldots w_n)$ of keywordsets, such that any consecutive two keywordsets have a common sub-keywordset. Their corresponding simplices share a face, and a "path" can be created from one P-cluster to another (cf. Section V L4 and L6). Two documents are clustered together if both documents have a keywordset that belongs to the same global C-concepts. The documents have addressed the I-concepts of this C-concept.

For further clarification, these concepts will be revisited in the discussion of the experimental result.

VI. Experimental Result

The documents used in this experiment are the collection of NSF Research Awards Abstracts which can be downloaded from the UC Irvine KDD Archive.[1] The number of documents used in the experimentation is 19,876 out of 129,000 documents. The documents in this data set consist of abstracts of all research that received awards from National Science Foundation (NSF) from the years of 1990 to 2003.

Tables I , II, and III show the initial result of this experiment. Note that the tables only show partial results from the whole result, since the whole result is too large to be displayed in one table.

Column A in the tables uniquely identifies the P-concept defined by the current tuple. Column B contains the relative cluster number to which this P-concept belongs. Column C states the number of documents in the data set that contain this P-cluster. The remaining numbered columns uniquely identify tokens. The high dimensional clusters are collected for clarity. Even though the result shows 7-keywordset as the maximum keywordset, it is easy to see now they can generalize n-keywordsets and build the mathematical structure. This alows one to capture the IDEA behind this set of documents.

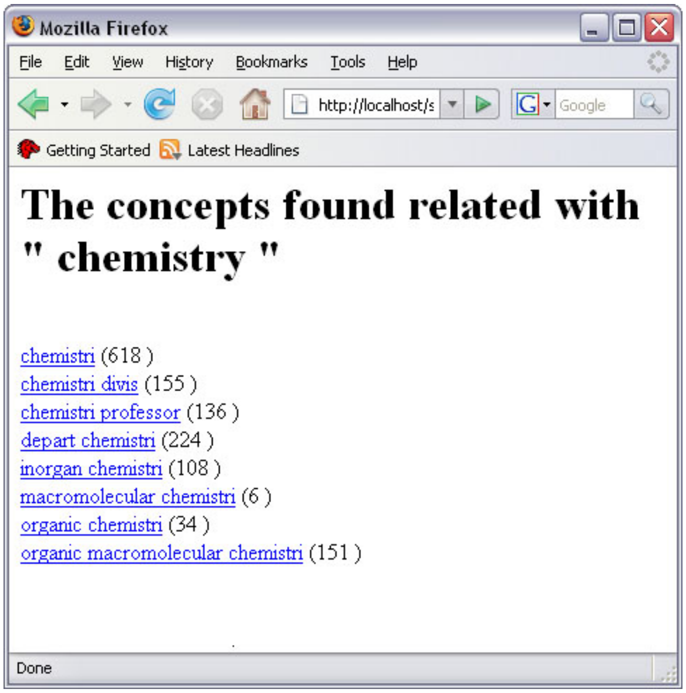
A. Clustering by Concepts

Taking a closer look at Table I. It represents an interesting subcomplex of the KSC (see Section V) produced from the NSF document set. Each tuple in Table I represents a cluster, called a P-cluster. P-concepts are used for clustering (see Section V). Column A enumerates P-clusters. Column C indicates the number of documents in this P-cluster. The remaining columns list the keywords in this P-concept. Table I shows two C-concept clusters, the sub-complex that consists of the 2-simplex $\Delta(earth, miner, seismolog)$ representing a relative cluster. If dropped, one can make the two C-concept clusters disjoint.

In traditional clustering, a document set is partitioned into disjoint groups, namely, equivalence classes of documents. However, many documents are inter related in some concepts yet completely unrelated in others. A concept-based clustering where using the conceptual structure of IDEA to group the concepts is proposed.

Figure 2 shows the screenshot of a demo program that retrieves the n-keywordsets in response to a query "chemistry." The program returns the P-concept which contains "chemistry." Please note all the keywords shown in P-concept have been stemmed. The numbers in parentheses are the number of documents containing the P-concept. The picture shows some concept related with "chemistry", which can be "chemistri division", "chemistry professor", "inorganic chemistry", "organic chemistry", "macromolecular chemistry" or "organic macromolecular chemistry".

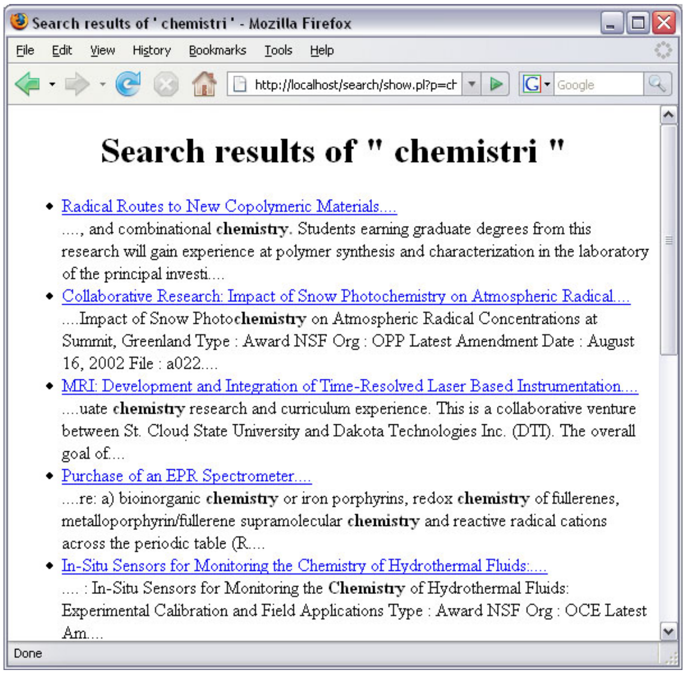
Figure 3 shows all the documents that are clustered under the concept of "chemistri." The documents shown under the cluster of "chemistri" are documents that contains "chemistri", but not the superset ("chemistri divis", "chemistri professor", etc). Likewise, the documents that are clustered by "chemistri divis" are documents that contain "chemistri divis" but not the subset ("chemistri" or "divis"). In other words, the documents shown are documents that contain maximal simplices.



The concepts found related with " chemistry "

chemistri (618)
chemistri divis (155)
chemistri professor (136)
depart chemistri (224)
inorgan chemistri (108)
macromolecular chemistri (6)
organic chemistri (34)
organic macromolecular chemistri (151)

Fig. 2. Retrieval of P-concept


Search results of ' chemistri ' - Mozilla Firefox
File
Edit
View
History
Bookmarks
Tools
Help
http://localhost/search/show.pl?p=ch
Google
Done

Search results of " chemistri "

- Radical Routes to New Copolymeric Materials....
  ...., and combinational **chemistry**. Students earning graduate degrees from this research will gain experience at polymer synthesis and characterization in the laboratory of the principal investi....
- Collaborative Research: Impact of Snow Photochemistry on Atmospheric Radical....
  ....Impact of Snow Photo**chemistry** on Atmospheric Radical Concentrations at Summit, Greenland Type : Award NSF Org : OPP Latest Amendment Date : August 16, 2002 File : a022....
- MRI: Development and Integration of Time-Resolved Laser Based Instrumentation....
  ....uate **chemistry** research and curriculum experience. This is a collaborative venture between St. Cloud State University and Dakota Technologies Inc. (DTI). The overall goal of....
- Purchase of an EPR Spectrometer....
  ....re: a) bioinorganic **chemistry** or iron porphyrins, redox **chemistry** of fullerenes, metalloporphyrin/fullerene supramolecular **chemistry** and reactive radical cations across the periodic table (R....
- In-Situ Sensors for Monitoring the Chemistry of Hydrothermal Fluids:....
  .... : In-Situ Sensors for Monitoring the **Chemistry** of Hydrothermal Fluids: Experimental Calibration and Field Applications Type : Award NSF Org : OCE Latest Am....

Fig. 3. Documents Clustered by P-Concept of "Chemistri"

[1] http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.html

A	B	C	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1	7	radar	gyroscop	earth	satellit	receiv	pin	point						
2	1	12	radar	gyroscop	earth	satellit	receiv	pin							
3	1	19	radar	gyroscop		sattelit	receiv								
4	1	19	radar	gyroscop		satellit	receiv								
5	1	12			earth	satellit	receiv								
6	1	14			earth	satellit									
7	-	7			earth					miner					seismolog
8	2	6									radiogen	tracer	isotop	geochemistri	seismolog
9	2	6									radiogen	tracer	isotop		seismolog
10	2	6										tracer	isotop	geochemistri	seismolog
11	2	6									radiogen	tracer	isotop	geochemistri	
12	2	6									radiogen	tracer	isotop		
13	2	24									radiogen		isotop		

TABLE I
RESULT SAMPLE 1

A	B	C	14	15	16	17	18	19	20	21	22	23	24	25
8	2	6	magnetotellur	sound										
9	2	6	magnetotellur	sound										
10	2	6	magnetotellur											
11	2	6												
12	2	6												
13	2	24												
14	-	7		sound	ice									
15	3	7			ice	upper	ocean	primari	product					
16	3	10					ocean	primari	product					
17	3	8			ice			primari	product					
18	3	7			ice		ocean	primari						
19	3	12				upper	ocean							
20	-	25			ice		ocean							
21	4	6					ocean			atmospher	variabl	ecosystem	respons	ross
22	4	7					ocean			atmospher		ecosystem	respons	
23	4	8					ocean			atmospher		ecosystem		
24	4	12					ocean					ecosystem		
25	-	4										ecosystem	respons	
26	5	5											respons	
27	5	5											respons	
28	5	5											respons	
29	5	5												

TABLE II
RESULT SAMPLE 2

VII. DOCUMENT INDEX

The documents index is built based on the n-keywordsets generated by Association Rule Mining process. These n-keywordsets represent simplices, and each simplex represents a concept within documents. Each simplex consists of some documents. In other words, documents are grouped by simplices.

The index takes the form of inverted lists,where each simplex points to list of documents where the simplex occurs. The index representation of these simplices is stored in a relation of $< simplex\_id, prefix\_id, vertex, dimension >$ where each tuple is an n-simplex with *simplex_id*. The value of n is denoted by *dimension* field. The field *vertex* is a term in an n-simplex, with *prefix_id* pointing to another vertex which is the prefix of the current vertex. For example the following simplices

- (wall, street, journal)
- (wall, street)
- (wall)

can be represented as the following relation :

simplex_id	**prefix_id**	**vertex**	**dim**
1	0	wall	1
2	0	street	1
3	0	journal	1
4	1	street	2
5	4	journal	3

The individual term *wall* , *street* and *journal* have no other terms as their prefix, since they are single terms which is denoted by their *dimension*. The term *street* with simplex id 4 has simplex id 1 as its prefix. The dimension of its simplex is 2, as denoted by its dimension. Simplex id 4 represent the term "*wall street*". Likewise for simplex id 5, which is a 3-keywordsets, representing "*wall street journal*".

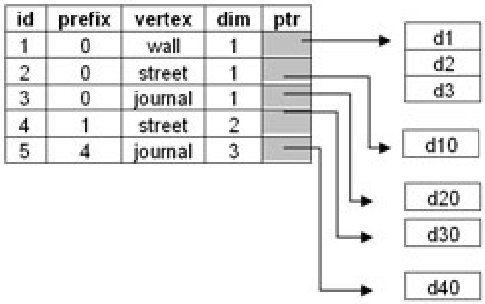
By representing simplices in such relation, space can be saved by "compressing" the length of n-simplex. A simplex that has prefix_id = 0 is 1-simplex, and a simplex_id which is not referenced in the prefix_id column anywhere in the relation, is a maximal simplex. Each tuple in the index point to list of documents where the terms occurs. The index can be used to respond to the user's query and retrieve simplices which, in turn, retrieve documents grouped by the simplices.

In the actual implementation, the order of terms is not important.

A	B	C	27	28	29	30	31	32	33	34	35
12	2	6									
13	2	24									
14	-	7									
15	3	7									
16	3	10									
17	3	8									
18	3	7									
19	3	12									
20	-	25									
21	4	6									
22	4	7									
23	4	8									
24	4	12									
25	-	4	marin	polar	climat						
26	5	5			climat	milankovitch	forc	proxi	seri	obtain	
27	5	5			climat	milankovitch	forc	proxi			
28	5	5			climat	milankovitch	forc				
29	5	6			climat	milankovitch	forc				
30	5	8				milankovitch	forc				

TABLE III

Result Sample 3



id	prefix	vertex	dim	ptr
1	0	wall	1	
2	0	street	1	
3	0	journal	1	
4	1	street	2	
5	4	journal	3	

Fig. 4. Index of simplices

The terms stored in the index is ordered alphabetically for efficiency purposes, therefore the term “*wall street*” in the index might be represented as “*wall*” that has “*street*” as its prefix if the terms are sorted in increasing order. In the explanation above, the terms order is shown such that it is easier to grasp the idea.

VIII. Conclusion

The idea in this project is new, and experiments are still in progress. Each document may have several concepts. Therefore, many documents may be intertwined among themselves. Consequently, cleanly separating documents may not always be possible. One interesting application of the approach described in this project is that two documents can be detected to have the same concept just by comparing the simplicial complex structure. Let A and B be two document sets written in different language, where B is a translation of A then the simplicial complex structure of A and the simplicial complex structure of B are isomorphic. Using this theorem, two sets of documents written in different languages can be identified similar even without translation.

Finally, it is important to note that the coincidence of the closed condition and the Apriori Condition seems to indicate that the simplicial complex is the natural structure to carry the semantics of associations. In this paper, only the SUPPORT value of Association Rule is used for high dimension simplex. The reason for this choice is computational. Another possibility is to use high dimensional TFIDF. This may be able to remove some high dimension stop words.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th VLDB Conference*, 1994.

[2] 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

[3] 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.

[4] 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.

[5] T. Y. Lin, I-Jen Chiang, ”Granulate and Conquer: Clustering Web Pages Semantically using Combinatorial Topology” In: the 10th Conference on Artificial Intelligence and Applications,” Dec 2-3, 2006, Kaohsiung, Taiwan.

[6] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[7] G. Salton and C. Buckley. Term Weighting Approaches in Automatic Text Retrieval, 1960, Information Processing and Management, 24, Vol 5, 513-523

[8] E. Spanier. *Algebraic Topology*. McGraw-Hill Book Company, New York, NY, 1966.

[9] A. Sutojo. *Concept-Based Document Index System*. Master Thesis, San Jose State University, 2006.