Experimental building automatic Thesaurus By Using Data Mining

Dr.Hadeel Sh.Al-Obiady* Dr.Arwa I.Al-Yasiri*

ABSTRACT

In this paper, a new approach was presented, this approach building automatic thesaurus in the software engineering subject.

The proposed approach depend on the data Mining (that refers to the overall process of discovering patterns or building models from a given data set) for indexing all the terms that founded into abstracts of the university thesis , this done by:

●Association rule: - this algorithm was used into data mining to find large item set and it is as a tool of discovery rule in our approach we used it to compute the frequency of the words into text and determined the keywords.

●Clustering techniques: it's used to classification of patterns

(observation, data items, or feature vectors into groups).

This new approach introduce automatic thesaurus that can be used as an effective tools for information retrieval.

Keywords:- Data mining, automatic thesaurus , Association rule .

***Computer Science & Information System Department/**
Al- Mansour University College

1. Introduction:-

The Indexing is mentality operation that depends on the indexer capabilities and his potentialities to pick up the concepts that mentioned in documents than converted it to terms that indicating to these concepts.

Many of researchers are starting since middle of twentieth century attempted to achieve this operation automatically, especially after increased the literature that published into internet and the big and rapid development in the hardware and software.

Several experiments in automatic generation of thesaurus have been carried out in which relationship between terms have been determined by taking into account the number of documents in which the respective terms occur jointly. Various clustering techniques have been investigated out of a range of similarity criteria. The role played by similarity criteria in obtaining the environment of each term and the use of this environment for retrieval has been explored.

Computational procedures for generating thesaurus include keyword statistical , calculation of Tanimoto coefficient, matrix inversion , formation of similarity matrix , automatic cluster analysis using minimal tree procedure and compilation of groups and main groups of descriptors [8].

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses.

Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by

retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve.

They scours databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations .

In this paper we suggest a new approach to build automatic thesaurus this approach based on deduction by using data mining techniques such as Association rule and clustering terms [13, 14].

2. Subject Indexing:-

Indexing has traditionally been one of the most important research topics in information science. Indexes facilitate retrieval of information in both traditional manual systems and newer computerized systems. Without proper indexing and indexes, search and retrieval are virtually impossible[16].

Observes that in library science, indexing records the values of various attributes expected to be used as a basis for searching. Simply put, the goal of subject indexing is to produce a set of attributes that represent the content or topics of a document.

Traditionally, a great deal of effort has been invested in subject indexing. Traditional human indexing has two main tasks .The first is to recognize and select the essence, or "about-ness," of a text. This is done by reading or scanning the document. The second task is to represent the essence of the text. In this process, the indexer assigns a set of index terms

to represent the central topics of the document.

Ideally, an indexer reads the full text of a document before determining the "about-ness" of it. However, indexers typically work under severe time constraints. For example, determined an optimum indexing time of four minutes, which makes reading the full text impossible in most cases. As a result, most indexers scan a document for its main topics [1]. While scanning, indexers normally engage both perceptual and conceptual faculties. Perceptual processes employ information based on the actual content of the document. Conceptual processes, on the other hand, use global knowledge not contained in the document itself, but rather in the knowledge that the author implies in the document or in the domain knowledge the indexer possesses. After determining the main concepts of a document, an indexer selects a set of conceptual terms to represent it. However, it is difficult to select a small set of "best" terms among all the possible terms that can represent a document [5, 8] .

3. Automatic indexing-

It is a process of assigning and arranging index terms for natural-language texts without human intervention. For several decades, there have been many attempts to create such processes, driven both by the intellectual challenge and by the desire to significantly reduce the time and cost of producing indexes.

Dozens if not hundreds of computer programs have been written to identify the words in a text and their location, and to alphabetize the words. Typically, definite and indefinite articles, prepositions and other words on a so-called stop list are not included in the program's output. Even some word processors provide this capability. Although automated indexing is a pipe dream, computers are absolutely essential in creating all but the simplest indexes. Most indexers would not do their job without indexing software [12]

There are two methods to generated automatic thesaurus that is a statistical method and linguistic method, Lancaster and Warner summarized the comparison between Traditional technique and other techniques in building of thesaurus: [18].

Traditional		Automatic	
		Statistical	linguistic
Selection Procedure	Verification of terms by selection of terms committee	Selected keyword depend on frequency	Selected keyword/ Expressions by linguistic analysis (grammars or morphology analysis)
Organization procedure	Verifying, from semantic relational	Grouping frequency keyword depend on associated with together or ratio frequency	finding out semantic relational from text
Expression format	List of terms and semantic relational	Clustering group for keyword or expressions	Semantic network by terms and their relational

Table (1) the tunes of indexing

4.Data Mining (DM):-

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on most important information in their data warehouses. [6].

Data mining is the act of drilling through huge volumes of data in order to discover relationships, or to answer specific questions that are too broad in nature for traditional query tools [10].

Data mining is traditional data analysis methodology updated with the most advanced analysis techniques applied to discovering previously unknown patterns [9].

Data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules [11].

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following Figure (1) shows data mining as a step in an iterative knowledge discovery process [2, 14].

The Knowledge Discovery in Databases process comprises a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

- **Data cleaning: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.**
- **Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.**
- **Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.**
- **Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.**
- **Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.**
- **Pattern evaluation: in this step, strictly interesting patterns**
  representing knowledge are identified based on given measures.
- **Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.**

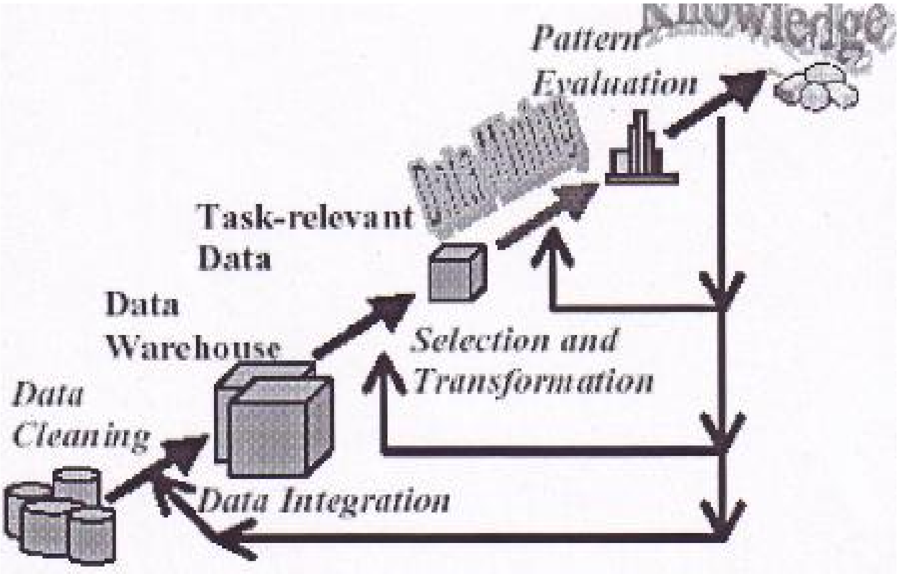


Figure (1) :Data Mining as a Step in an Iterative Knowledge Discovery Process

It is common to combine some of these steps together. For instance, *data cleaning* and *data integration* can be performed together as a preprocessing phase to generate a data warehouse. *Data selection* and *data transformation* can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data [1, 3].

5. Association Rule:-

Association rules are one of the promising aspects of data mining as knowledge discovery tool, and have been widely explored to data. They allow capturing all possible rules that explain the presence of some attributes according to the presence of other attributes. An association rule, X => Y , is a statement of the form "for a specified fraction of transaction , a particular value of an attributes set X determines the value of attributes set Y as another particular value under a certain confidence".

Thus association rule aim at discovering the patterns of co occurrence of attributes in a database [3, 4].

5.1. Formal Definitions:-

The formal definition of association rule is the following [3,4]: Let $G = \{i_1, i_2, ..., i_m\}$ be a set of literal, called items. Let D be a set of transactions, where each transaction T is a set of items such that T Í G Associated with each transaction is a unique identifier, called its TID.

Definition (1): **-An item X is a set of items in I , An item set X is called K-item set if it contains K items from I.**

Definition (2):- **A transaction T satisfies an item set X if X Í T . The support of an item set X in D, supports $_D$(X), and is the number of transactions in D that satisfies X.**

Deflnition(3):- **An item X is called Large item set if support of X in D exceeds a minimum support threshold explicitly declared by the user and a small item set otherwise.**

Definition (4):-

An association rule is an implication of the form X Þ Y, where X Ì G, Y Ì G, and X intersection Y = F . The support and confidence of an association rule (X Þ Y) are calculated by the following two equations:

The rule X=>Y holds in the transaction set D with confidence c Where c = support D(XUY)/ support D(X).The rule X=>Y has support s in D if the function s of the transactions in D contain X U Y.

$$\text{Support}=\frac{\textit{The Number of Transaction Contain XandY}}{\textit{Total Number of Transaction}}$$

If its support and confidence are equal to or greater than the user specified values. The goal of association rules is to find the relationship between any combinations of items.

Example 0):- Consider the example transaction database ETDB in table (3). there are six transaction in the database with Transaction IDentifiers, TIDs ,1,2,3,4,5, and 6 . the set of item sets I=jA,I3,C,D,E,Fj, each item is an abbreviation of book title in bookshop sales as shown table (2) . There are totally (2^6-1)=63 nonempty item set (each non-empty subset of I is an item set).{A} is a 1- item set and {A,13) is a 2-items set and so on[3,4].

Support (A) =3 since three transactions include A in it. Let us assume that the minimum support (minsup) is two (approximately taken as 33%). Then {A,B,C,D,E,AB AC AE BC D,BD, BE, CT ,CE, DE, ABC, ABE, ACE, BCD, BCE ,BDE,CDE, ABCE, BCDE } are the set of large item set ;since their support is grater than or equal to 2.(33% * 6) , and the remaining ones are small item sets, there are two item sets , ABCE and BCDE called maximal item sets ; all other large item set are subset s of one of them . Table (4) depicts large item set with their support. Let's assume that the, minimum confidence (minconf) is set to 100% , then A=> B as an association rule with respect to the specified minsup and minconf (its support is 3) , and its confidence is :-

$$\frac{\text{Support ETDB(AD)}}{\text{Support ETDB(A)}} \times 100 = 3/3 \times 100 = 100\%$$

On other hand, the rule B=>A is not valid association rule since its confidence is 50%. The table (5) depicts the association rules that be mined from database ETDB according to 100% confidence and 33% minsup value.

Item	Book Title
A	From Here to Eternity
B	Love at the Time of Cholera
C	Gone with the wind
D	The Moon and the Fences
E	The tree and assassination of Marzooq
F	The monster

Table (2) the items abbreviations of database ETDB

Transaction TID	Item-(Books)
1	B, C, E
2	B,C, D, E
3	A, B, C, D, E
4	B, C, D
5	A, B, F
6	A, B, C, E

Table (3) A Transaction Data Base

Support	Item set	No.
6=100%	B	1
5=83%	C ,BC	2
4=67%	E, BE ,CE ,BCE	3
3=50%	A, D ,AB ,BD ,CD, BCD	6
2=33%	AC, AE, DE, ABC, ABE ,ACE, BDE, CDE, ABCE,BCDE	10

Table (4) Lar^{g}e item set with minsup 33%=2

A Þ B(3/3)	AC Þ B(2/2)	AC Þ BE(2/2)
C Þ B(5/5)	AE Þ B(2/2)	AE Þ BC(2/2)
D Þ B(3/3)	AC Þ E(2/2)	DE Þ BC(2/2)
E Þ B(3/3)	AE Þ C(2/2)	ABC Þ B(2/2)
D Þ C(4/4)	DE Þ B(2/2)	A]BE Þ C(2/2)
E Þ C(4/4)	DE Þ C(2/2)	CE Þ B(2/2)
ABE Þ C(2/2)	ACE Þ B(2/2)	ABC Þ E(2/2)

Table (5) Associations rules with minconf--100%

Rule Generation Rule

Step 1: for all large k-items l_k ,K >= 2, in L do
Step 2: begin
Step 3 -H_i =[consequents of rules from l_k with one item Step 4: in the consequent]
Step 5 :ap-gerules(l_k, H
Step 6:end.

ap-gerules($l_{k,}$ H_m) Algorithm

```
Step 1: if k> m+1 then
Step 2: begin
Step 3: $H_{m+l}$ =appriori-gen($H_m$)
Step 3: for all $h_m+_l$ to $H_{m+l}$  do
Step 4: begin
Step 5:conf= support D(Lk)/ support D(Lk, $h_{m+l}$    )
Step 6: if conf>= minconf then
                Add (Lk - $h_{m+l}$) => $h_{m+l}$ to the rule set
Step 7: else
        Delete $h_{m+l}$ from $H_{m+l}$
Step 8: end
Step : ap-gerules(lk, $H_{m+l}$   )
Step 9: end.
```

The candidate generation algorithm

apriori _ gen(L_{k-1})

```
Step 1: ck= f
Step 2: for all item set X  I E Lk.I and Y  I  $L_{k-1}$ do3
Step 3: if X1 = Y1 U  U $X_{k-2}$ = $Y_{k-2}$ U $X_{k-1}$ <= $Y_{k-1}$ then
begin    C'=XIX2 $X_{k-1}$ $Y_{k-1}$
Step 4: Add C to Ck
Step 5: End
step 6: Delete candidate item sets in Ck whose any subset is not
in $L_{k-1}$
```

Th e apriori algorithm

```
Step 1: L1= {   largel- item set   }
Step 2: K=2
  Step 3: While $L_{k - 1}$ =O do
Step 4 :Begin

Step 5: $C_k$=apriori- gen ($L_{k - 1}$ )
Step 6: For all transctionst in D do
Step 7:Begin
Step 8: C=subset($C_{k - 1}$)
Step 9 : For all candidate c I  C do
Step 10 : c. count =c. count+.1
Step 11 : end
Step12:$L_{k=}${ c I C' \ c. count >=minsup }
Step 13 :  K=k+1
Step 14: End
```

6. Clusterinx

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis [9]. From a machine learning perspective clusters correspond to *hidden patterns*, the search for clusters is *unsupervised learning*, and the resulting system represents a *data concept* .

From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, marketing, medical diagnostics, computational biology, and many others.

Clustering is the subject of active research in several fields such as statistics, pattern recognition, and machine learning. This survey focuses on clustering in data mining. Data mining adds to clustering the complications of very large data sets with very many attributes of different types. This imposes unique computational requirements on relevant clustering algorithms [7].

6.1 What is Clustering in Data Minim?

Clustering is a process of partitioning a set of data (or objects) in a set of meaningful sub-classes called clusters. It is helps users understand the natural grouping or structure in a data set.

Cluster is a collection of data objects that are "similar" to one another and thus can be treated collectively as one group. Clustering is unsupervised classification (no predefined classes).

Supervised Classification = Classification

(We know the class labels and the number of classes).

Unsupervised Classification = Clusterin

(We do not know the class labels and may not know the number of classes).

Clustering is the process of grouping physical or abstract objects into classes of similar objects, it is defined as [7, 15]: -

1. A cluster is a subset of records, which are "similar".
2. A subset of records such that the distance between any two records in the cluster is less than the distance between any record in the cluster and any record not in it.
3. A connected region of a multidimensional space containing a relatively high density of records.
4. Cluster is a collection of data objects, these are
   - ♦ Similar to one another within the same cluster.
   - ♦ Dissimilar to the objects in other clusters.
5. Clustering is unsupervised classification: no predefined classes
6. Cluster analysis is grouping a set of data objects into clusters.

Figure (2): the cluster operation

7. The Proposed Thesaurus:-

The **determine to methodology of thesaurus setting is important and hared operation , this operation need from developer the scientific and conceptual analysis and he must put it into standard levels that used to build automatic thesaurus.**

7.1 *The methodolo,-v of thesaurus setting:-*

The proposal thesaurus was work with the following issues:-

1. ***Thesaurus lan^{a}ua^{a}e:-*** ***will*** **be English language (uni-language) , then the descriptors taken from natural English language terms and the indexing language (documentation language) will be consider by extraction approach this means the terms from abstracts thesis that publishing in scientific journal and university thesis , so this language is free documentation language.**

2. ***Subiect space of the thesaurus: -*** **the proposal thesaurus is small thesaurus and it have (100-500) important descriptors and it is useful for indexing in the narrow space, this lead to characterized it accuracy of its terms.**

3. ***specify the shape of thesaurus :-*** **the proposal thesaurus consist of introduction identification of it and through it we can known about subject space and artist characteristics , how to build it , and using it to save and information retrieval .**

7.2 The thesaurus Design:--

The proposed work suggest a system that design and build automatic thesaurus into software Engineering using data mining techniques ; this system depend in its work into Association Rule to find Large Item sets and consider these item sets as a main features to clustering items later.

Firstly this system accept theses abstracts and build a Data Base of them and finally cluster interest items. The general algorithm is:-

The General Algorithm

Input: - abstracts thesis of software engineering (Text files (*.txt))
Output: - tables that contain clustered items

```
Step 1: Begin
Step 2: Build a Data Base that contains the following tables:
        1. table abstract
        2. table abbreviations
        3. table_ belongs terms
Step 3: using Lexical analysis to eliminate stop word list
Step 4: set minimum support= 2;
Step 5: for each word find the frequency of it
Step 6:- build a table that contain the words that
         greater or equal than minor value.
Step 7: call combinations algorithm to generate
Token/Word table
Step 8: call Association -Rule algorithm to find lager item
set
Step 9: using these item sets to cluster items into narrow,related
and broad terms
Step 10: Display thesaurus
Step 11: end.
```

The combination algorithm

Input: - table of word

Output: - table of Phrase / word

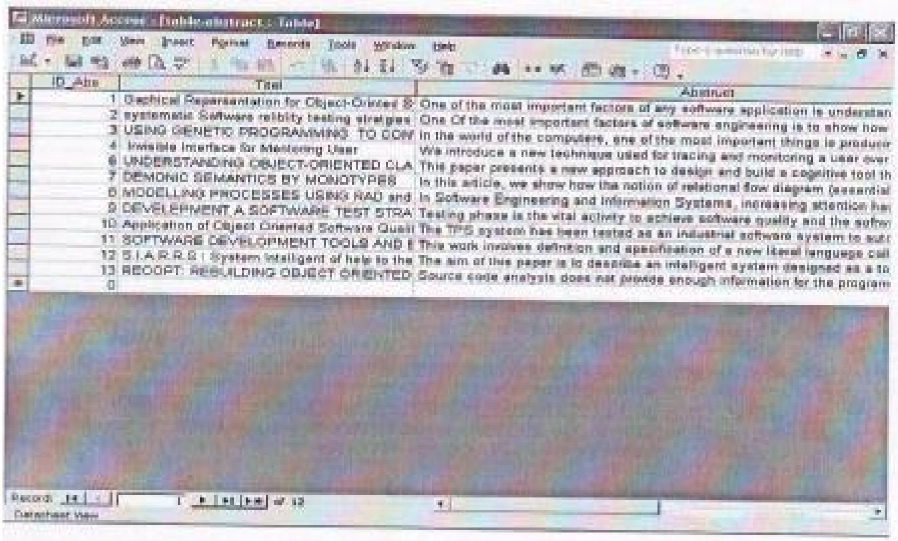
```
Step 1:- Begin
Step 2:- i=0;
Step 3:- count= the no. of
record of table word Step
4:- while i<= count do
          Begin
         Phrase =word $_{(i)}$+ word (i+1);

          Search Phrase into table- word
         If found then
                    Begin
                         Find its frequency
                          Add it into  table Phrase / word
                          Combination (Phrase, I)
                        End
               Else
                    If i=count then add word into table token
                    I=i+l;
Step 5:- End;
Step6:- End.
```

7.4 *The Implementation system:-*

First: **the proposal system was first establishing a data base that contains these tables:**

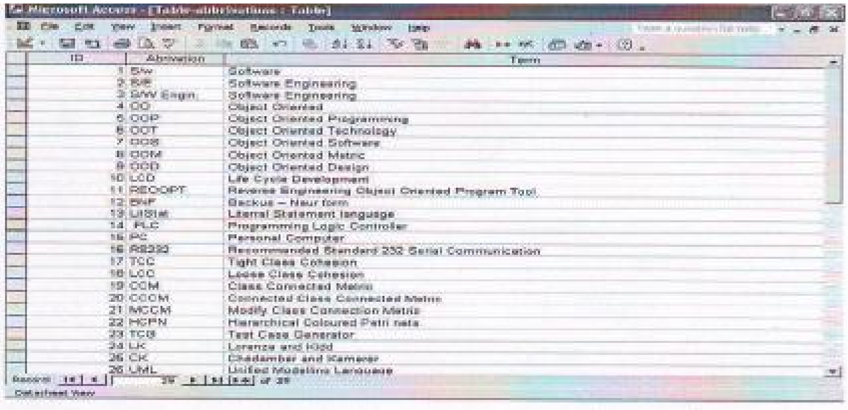
1. *Table(1) (table-abstract):-*

This table contains all theses abstracts (ID, Title, abstract)



ID_Abs	Titel	Abstract
1	Graphical Represantation for Object-Oriented S	One of the most important factors of any software application is understan
2	systematic Software relibity testing stratgies	One Of the most important factors of software engineering is to show how
3	USING GENETIC PROGRAMMING TO COM	in the world of the computers, one of the most important things is producir
4	Invisible Interface for Monitoring User	We introduce a new technique used for tracing and monitoring a user over
6	UNDERSTANDING OBJECT-ORIENTED CLA	This paper presents a new approach to design and build a cognitive tool th
7	DEMONIC SEMANTICS BY MONOTYPES	In this article, we show how the notion of relational flow diagram (essentia
8	MODELLING PROCESSES USING RAD and	In Software Engineering and Information Systems, increasing attention ha
9	DEVELEPMENT A SOFTWARE TEST STRA	Testing phase is the vital activity to achieve software quality and the softw
10	Application of Object Oriented Software Quali	The TPS system has been tested as an industrial software system to aut
11	SOFTWARE DEVELOPMENT TOOLS AND E	This work involves definition and specification of a new literal language cal
12	S.I.A.R.R.S : System Intelligent of help to the	The aim of this paper is to describe an intelligent system designed as a to
13	REOOPT: REBUILDING OBJECT ORIENTED	Source code analysis does not provide enough information for the program
0		

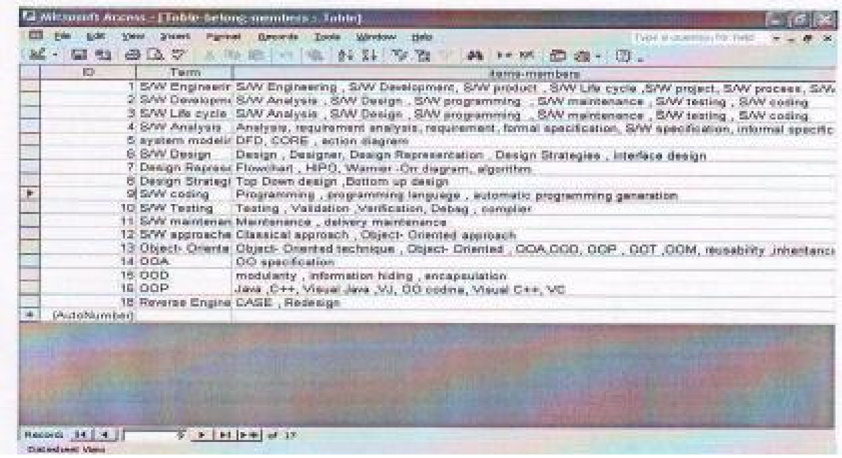
2. Table (2)(table-abbreviations):- **this table contain each term in the Software Engineering and it abbreviations (ID, Abbreviation, Term)**


Microsoft Access - [Table-abbriviations : Table]
File Edit View Insert Format Records Tools Window Help
Datasheet View

ID	Abbriviation	Term
1	S/w	Software
2	S/E	Software Engineering
3	S/W Engin.	Software Engineering
4	OO	Object Oriented
5	OOP	Object Oriented Programming
6	OOT	Object Oriented Technology
7	OOS	Object Oriented Software
8	OOM	Object Oriented Metric
9	OOD	Object Oriented Design
10	LCD	Life Cycle Development
11	REOOPT	Reverse Engineering Object Oriented Program Tool
12	BNF	Backus – Naur form
13	LitStat	Literal Statement language
14	PLC	Programming Logic Controller
15	PC	Personal Computer
16	RS232	Recommended Standard 232 Serial Communication
17	TCC	Tight Class Cohesion
18	LCC	Loose Class Cohesion
19	CCM	Class Connected Metric
20	CCCM	Connected Class Connected Metric
21	MCCM	Modify Class Connection Metric
22	HCPN	Hierarchical Coloured Petri nets
23	TCG	Test Case Generator
24	LK	Lorenze and Kidd
25	CK	Chadamber and Kamarer
26	UML	Unified Modeling Language

Record: 26 of 26

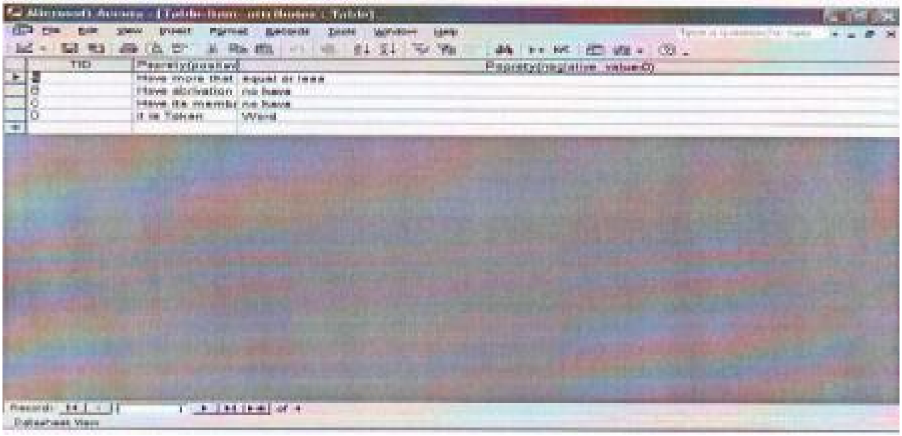
3. Table (3)Belong- items):- **this table contain each term have theme member terms that belong to it.**


Microsoft Access - [Table-belong-members : Table]
File Edit View Insert Format Records Tools Window Help
Type a question for help
Datasheet View

ID	Term	items-members
1	S/W Engineerir	S/W Engineering , S/W Development, S/W product , S/W Life cycle ,S/W project, S/W process, S/W
2	S/W Developm	S/W Analysis , S/W Design , S/W programming , S/W maintenance , S/W testing , S/W coding
3	S/W Life cycle	S/W Analysis , S/W Design , S/W programming , S/W maintenance , S/W testing , S/W coding
4	S/W Analysis	Analysis, requirement analysis, requirement , formal specification, S/W specification, informal specific
5	system modeli	DFD, CORE , action diagram
6	S/W Design	Design , Designer, Design Representation , Design Strategies , interface design
7	Design Represi	Flowchart , HIPO, Warnier -Orr diagram, algorithm
8	Design Strategi	Top Down design ,Bottom up design
9	S/W coding	Programming , programming language , automatic programming generation
10	S/W Testing	Testing , Validation ,Verification, Debag , compiler
11	S/W maintenan	Maintenance , delivery maintenance
12	S/W approache	Classical approach , Object- Oriented approach
13	Object- Orienta	Object- Oriented technique , Object- Oriented , OOA,OOD, OOP , OOT ,OOM, reusability ,inheritanc
14	OOA	OO specification
15	OOD	modularity , information hiding , encapsulation
16	OOP	Java ,C++, Visual Java ,VJ, OO coding, Visual C++, VC
18	Reverse Engine	CASE , Redesign
(AutoNumber)		

Record: 9 of 17

4. Table (4) (item- attributes):- **this table contains all attributes that may be major features of thesis abstract.**



Microsoft Access - [Table item-attributes : Table]

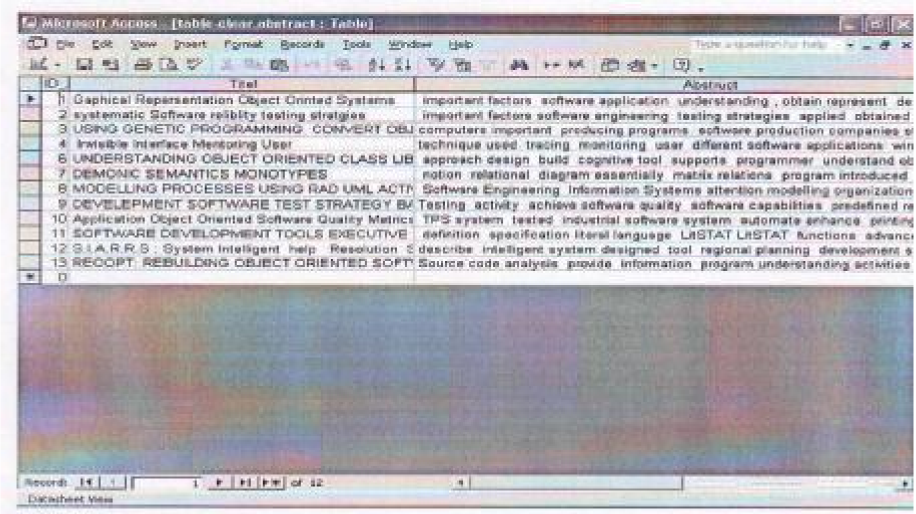
TID	Property(positive)	Property(negative_value0)
[illegible]	Have more that	equal or less
0	Have abrivation	no have
0	Have its memb	no have
0	it is Token	Word

Record: 1 of 4

Datasheet View

Second: - the proposal system was starting its processing steps:

1. ***Lexical analysis: -*** **in this step the text processing converts the text into a stream of tokens, including numbers, abbreviations, and alphanumeric sequences. There existing a large class of words that have no inherent meaning when taken out of context. For example,"a", "the", "of', "to" have no semantic meaning .Since these words to be among the most frequency occurring terms. So we filter them of text. This can be done by creating a list of terms, known as a stop word list, and generate the clear text form them this is shown in the table clear_ abstract.**

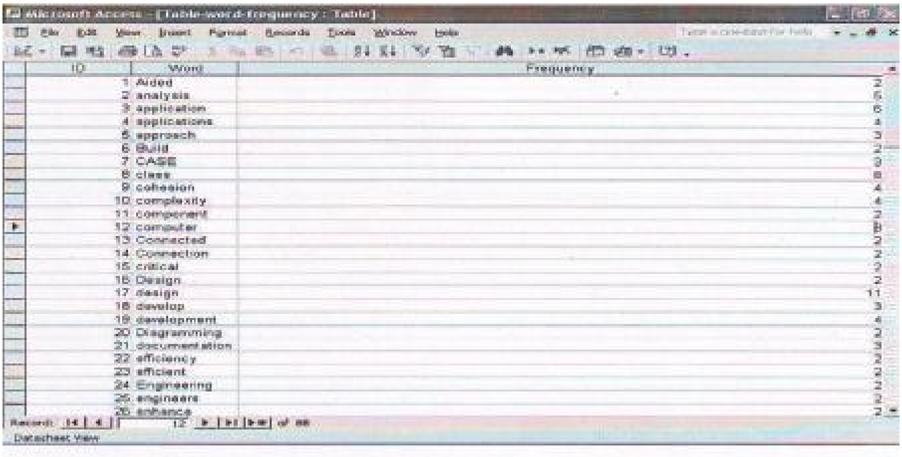


Microsoft Access - [table clear abstract : Table]
File Edit View Insert Format Records Tools Window Help
Type a question for help

ID	Titel	Abstract
1	Gaphical Repersentation Object Oriented Systems	important factors software application understanding , obtain represent de
2	systematic Software relibity testing stratgies	important factors software engineering testing strategies applied obtained
3	USING GENETIC PROGRAMMING CONVERT OBJ	computers important producing programs software production companies s
4	Invisible Interface Mentoring User	technique used tracing monitoring user different software applications win
6	UNDERSTANDING OBJECT ORIENTED CLASS LIB	approach design build cognitive tool supports programmer understand ob
7	DEMONIC SEMANTICS MONOTYPES	notion relational diagram essentially matrix relations program introduced
8	MODELLING PROCESSES USING RAD UML ACTI	Software Engineering Information Systems attention modelling organization
9	DEVELEPMENT SOFTWARE TEST STRATEGY BA	Testing activity achieve software quality software capabilities predefined re
10	Application Object Oriented Software Quality Metrics	TPS system tested industrial software system automate enhance printing
11	SOFTWARE DEVELOPMENT TOOLS EXECUTIVE	definition specification literal language LitSTAT LitSTAT functions advanc
12	S.I.A.R.R.S : System Intelligent help Resolution S	describe intelligent system designed tool regional planning development s
13	RECOPT REBUILDING OBJECT ORIENTED SOFT	Source code analysis provide information program understanding activities
0		

Record: 1 of 12

Datasheet View



Microsoft Access - [Table-word-frequency : Table]
File Edit View Insert Format Records Tools Window Help
Type a question for help

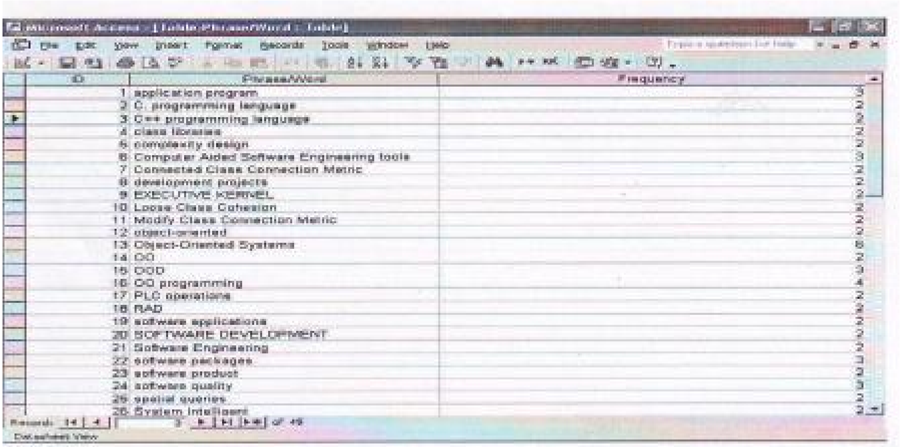
ID	Word	Frequency
1	Aided	2
2	analysis	6
3	application	6
4	applications	4
5	approach	3
6	Build	2
7	CASE	3
8	class	8
9	cohesion	4
10	complexity	4
11	component	2
12	computer	8
13	Connected	2
14	Connection	2
15	critical	2
16	Design	2
17	design	11
18	develop	3
19	development	4
20	Diagramming	2
21	documentation	3
22	efficiency	2
23	efficient	2
24	Engineering	2
25	engineers	2
26	enhance	2

Record: 12 of 88

Datasheet View

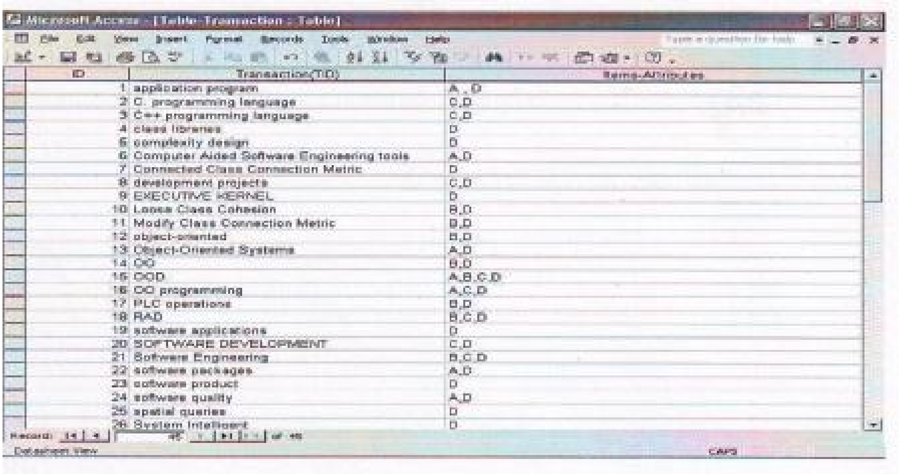
2. ***Compute the Frequency: -*** **In this step the proposal system was found the frequency for each token from abstract data base, and generates the Word- frequency table.**

3. ***Generate the Phrase /Word table:-*** in this step the proposal system was used the combination algorithm to generate the Phrase that occurred in our Data Base with its frequency and produce the Phrase/Word table .



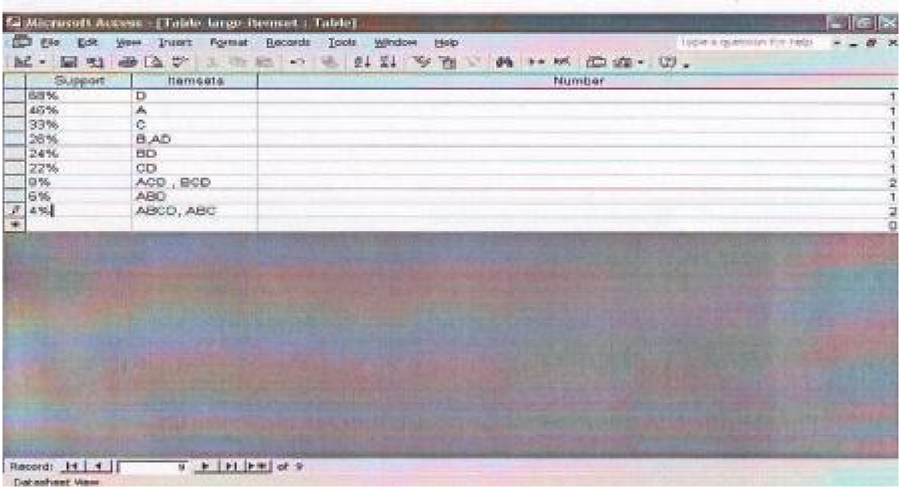
ID	Phrase/Word	Frequency
1	application program	3
2	C. programming language	2
3	C++ programming language	2
4	class libraries	2
5	complexity design	2
6	Computer Aided Software Engineering tools	3
7	Connected Class Connection Metric	2
8	development projects	2
9	EXECUTIVE KERNEL	2
10	Loose Class Cohesion	2
11	Modify Class Connection Metric	2
12	object-oriented	2
13	Object-Oriented Systems	6
14	OO	2
15	OOD	3
16	OO programming	4
17	PLC operations	2
18	RAD	2
19	software applications	2
20	SOFTWARE DEVELOPMENT	2
21	Software Engineering	2
22	software packages	3
23	software product	2
24	software quality	3
25	spatial queries	2
26	System intelligent	2

4. ***Genrate the Transactions (TID) :-*** in this step the proposal system generate the Transactions (TID) by used the attributes table and each transaction have some item -set and produce the Transaction table.


Microsoft Access - [Table-Transaction : Table]
File Edit View Insert Format Records Tools Window Help
Type a question for help
Record: 46 of 46
Datasheet View
CAPS

ID	Transaction(TID)	Items-Attributes
1	application program	A , D
2	C. programming language	C,D
3	C++ programming language	C,D
4	class libraries	D
5	complexity design	D
6	Computer Aided Software Engineering tools	A,D
7	Connected Class Connection Metric	D
8	development projects	C,D
9	EXECUTIVE KERNEL	D
10	Loose Class Cohesion	B,D
11	Modify Class Connection Metric	B,D
12	object-oriented	B,D
13	Object-Oriented Systems	A,D
14	OO	B,D
15	OOD	A,B,C,D
16	OO programming	A,C,D
17	PLC operations	B,D
18	RAD	B,C,D
19	software applications	D
20	SOFTWARE DEVELOPMENT	C,D
21	Software Engineering	B,C,D
22	software packages	A,D
23	software product	D
24	software quality	A,D
25	spatial queries	D
26	System Intelligent	D

5. ***Find large item —sets: -*** **in this step the proposal system using Association Rule to generate the large item sets and generate the large item-sets table.**


Microsoft Access - [Table large itemset : Table]
File Edit View Insert Format Records Tools Window Help
Type a question for help
Record: 9 of 9
Datasheet View

Support	Itemsets	Number
68%	D	1
46%	A	1
33%	C	1
26%	B,AD	1
24%	BD	1
22%	CD	1
9%	ACD , BCD	2
6%	ABD	1
4%	ABCD, ABC	2
		0

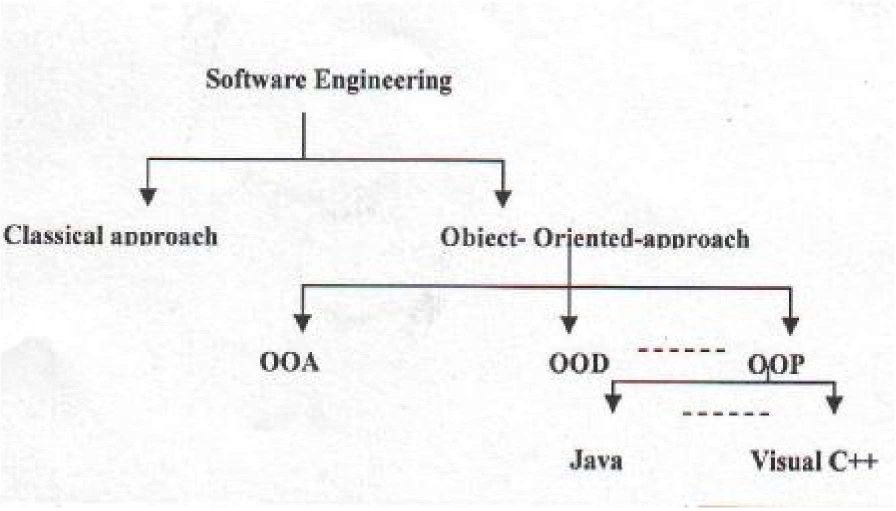
6. Clustering items :- in this step the proposal system using the lager item sets as main features to cluster items into (Broad Term (BT), Narrow Term (NT), Related Term (RT)) terms with minimum support value(miner value) equal 22%= 2 .For example

Software Engineering= ABCD, **this means Software Engineering is have frequency value more than two , is have abbreviations ,is have members terms , and it is a phrase , so it is a Broad Term (BT),this operation was done to cluster of its members.**

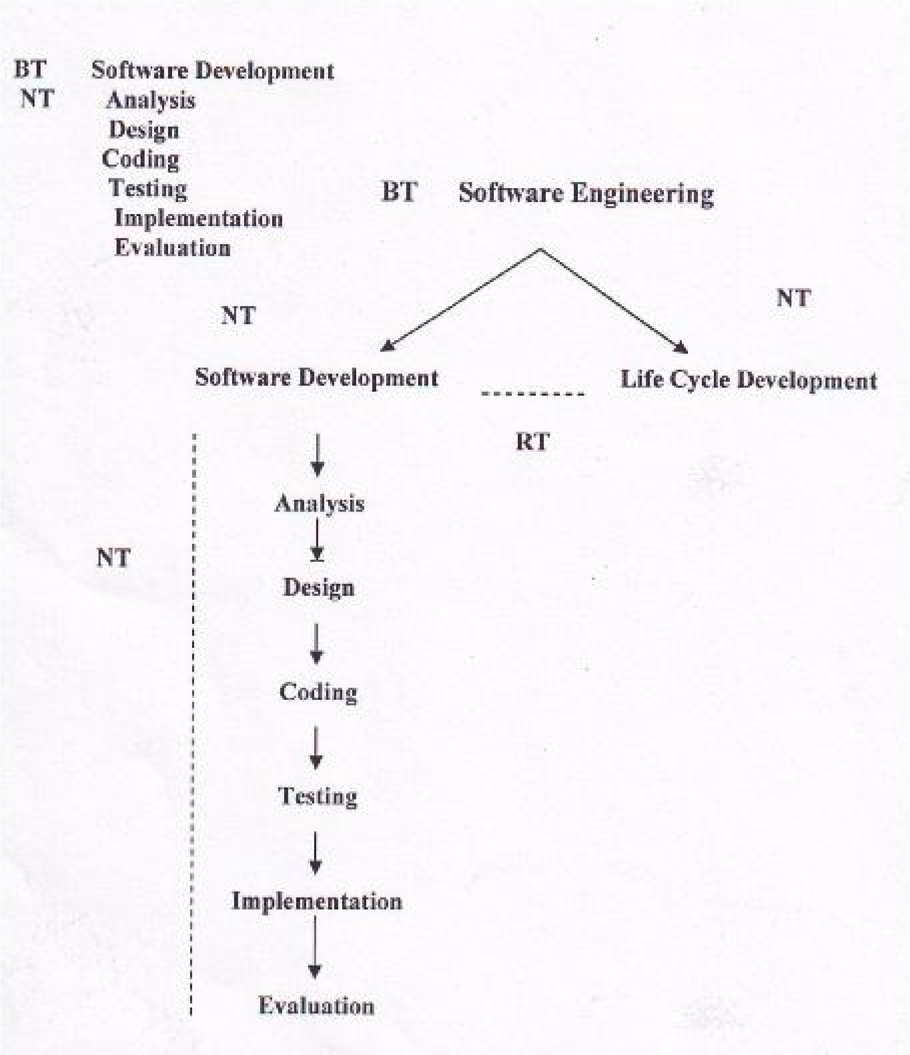
Ex 1:

BT	**Software Engineering**
NT	**Software Development**
RT	**Life Cycle Development**

Ex2:



Ex.3



BT Software Development
NT Analysis
Design
Coding
Testing
Implementation
Evaluation

8. Conclusions:-,

1. **This new approach introduce automatic thesaurus that can be used as an effective tools for information retrieval based on quickly and accurate that doesn't care about size of thesaurus.**
2. **This work can be used to find Keywords for any data entry.**
3. **Through this work we can use it to build a Web thesaurus from web link structure.**
4. **Although much effort has been devoted to hand — coded thesaurus, by this work developer reduce the effort by using automatic thesaurus using Association Rule to keep up with the speed of growth for new terms and concepts.**
5. **In the statistical approach the developer of thesaurus cannot deduction the main subject if it didn't occurred explicitly into text and it compute its frequency but by using Data Mining technique the miner could mixed between the statistical and Data mining techniques to avoid this problem because the Data Mining based on inference and prediction manner in its work.**

9.Refrences

1. **Agrawal R., Imielinski T., and Swami S., Mining Association rules between sets of items in large databases, Proc. of the ACM SIGMOD Conference on Management of Data, Washington, DC, May 1993.**

2. **Agrawal R.,and Ramakrishnan Srikant, Fast algorithms for mining association rules , in proceeding of 20'h Intl, conf. on Vary Large Data Base (VLDB'94),pages 487-499, Santiago de Chile ,sptermber,1994.**

3. **Alaa H. Al-Hamami, abass F Kader ,Hussein K.Al-khefaji,"Desgin and Implementation of Genenrate of large Dense, or sparce Database to test Association rules Miners" (selected Teachers papers), Scientific journal of Fedration of Arab Scintific Research Council, 2002.**

4. **Alas H. Al-Hamami, abass F Kader ,Hussein K.A1-khefaji, "a new Approach for mine negative association rule", journal of Al-Rafiaden Uni. Coll,No, 10, 10,2002.**

5. **Chung, Yi-Ming, William M. Pottenger, Bruce R. Schatz. Automatic Subject Indexing Using an associative Neural Network. available at htip://www.canis.uitic.edu/**

6. **D. Zhang and F. Currim, Data Mining. Technical report, 1996.**

7. **DAVID M. ROCKS AND RAN DAI, " Sampling and Sub Sampling for Cluster Analysis in Data Mining:**

With Applications to Sky Survey Data, Center for Image Processing and Integrated Computing, University of California, 2003.

8. Devadason, F. J. Generation Of Thesaurus In Different Languages A Computer Based System (PDF) available at:http://portal.acm.org

9. Marco BOTTA , "Clustering Techniques ",Dipartimento di Informatics UniversitAdi Torino,www.di.unito.it/-botta/didattica /clustering.html,2003.

10. Michael J. A. Berry and Gordan S. Linoff, Mastering Data Mining. John Wiley & Sons, Inc, 2000.

11. P. Adriaan and D. Zanting. Data Mining. Addison-Wesley: Harlow, England, 1996.

12. Tulic, Martin. Automatic indexing available at:http:// www.anindexer.com

13. Two Crows Corporation. Introduction to Data Mining and knowledge

14. U. Fayyad, G. Piatetsky-Shapiro,P. Smyth, & R. Uthurusamy, Advance in Knowledge Discovery & Data Mining. Cambridge, MA (The AAAI Press/The MIT Press), 1996.

15. Wei Wang," Clustering", COMP 290-90, UNIVERSITY of NORTH CAROLINA at CHAPEL HILL, Fall 2003.

16. الهبائلي،حسين.المعالجة اللغوية للمعلومات .زغوان:مؤسسة التميمي للبحث العلمي والمعلومات،1995.ص360

17. الهبائلي مصدر سابق.ص85

18. لانكستر،أف.ولفرد،آمي جي.اساسيات استرجاع المعلومات ترجمة حشمت قاسم.ط3 (مزيدة ومنقحة).- الرياض:مكتبة الملك فهدالوطنية،1997 .ص360