APPENDIX A

Cross Ontology Dataset

MeSH - Wordnet Cross Ontology Results

| Anemia - Appendicitis | 0.03125 | 0 | 0 |
| Dementia - Atopic Dermatitis | 0.0625 | 0 | 0 |
| Bacterial Pneumonia - Malaria | 0.15625 | 0.113 | 0 |
| Osteoporosis - Patent Ductus Arteriosus | 0.15625 | 0.122 | 0 |
| Acquired Immunodeficiency Syndrome - Congenital Heart Defects | 0.0625 | 0.084 | 0 |
| Otitis Media - Infantile Colic | 0.15625 | 0 | 0 |
| Meningitis - Tricarpal Atresia | 0.03125 | 0.025 | 0.008333333 |
| Sinusitis - Mental Retardation | 0.03125 | 0 | 0 |
| Hyperlipidemia - Hyperkalemia | 0.15625 | 0.182 | 0 |
| Hyperthyroidism - Hyperthyroidism | 0.40625 | 0.387 | 0 |
| Sarcoidosis - Tuberculosis | 0.40625 | 0 | 0 |
| Asthma - Pneumonia | 0.375 | 0.07 | 0.0119 |
| Diabetic Nephropathy - Diabetes Mellitus | 0.5 | 0.205 | 0.018333333 |

| Lactose Intolerance - Irritable Bowel Syndrome | 0.46875 | 0.047 | 0.005666667 |
| Urinary Tract Infection - Pyelonephritis | 0.65625 | 0.03 | 0.01 |
| Neonatal Jaundice - Sepsis | 0.1875 | 0 | 0 |
| Sickle Cell Anemia - Iron Deficiency Anemia | 0.4375 | 0.14 | 0.011666667 |
| Psychology - Cognitive Science | 0.59375 | 0.25 | 0.008333333 |
| Adenovirus - Rotavirus | 0.4375 | 0.16 | 0.018666667 |
| Migraine - Headache | 0.71875 | 0.042 | 0 |
| Myocardial Ischemia - Myocardial Infarction | 0.75 | 0.47 | 0 |
| Hepatitis B - Hepatitis C | 0.5625 | 0.42 | 0.016 |
| Carcinoma - Neoplasm | 0.75 | 0.17 | 0.04 |
| Failure to Thrive - Malnutrition | 0.625 | 0.043 | 0.014333333 |
| Breast Feeding - Lactation | 0.84375 | 0 | 0 |
| Antibiotics - Antibacterial Agents | 0.9375 | 1 | 0.03 |
| Pain - Ache | 0.875 | 1 | 0.021666667 |
| Malnutrition - Nutritional Deficiency | 0.875 | 1 | 0.143333333 |
| Chicken Pox - Varicella | 0.96875 | 1 | 0.247666667 |
| Down Syndrome - Trisomy 21 | 0.875 | 1 | 0.146666667 |

| Human correlation | 0.697181868 | 0.552957306 |

*Wordnet terms with orange colour

Source: Intelligent Systems Laboratory Technical University of Crete, Greece
Website: http://www.intelligence.tuc.gr/similarity/datasets/CrossOntologyDataset.pdf
APPENDIX B
Data Mining Algorithms

PREFIXSPAN ALGORITHM

- **Input** of the algorithm: A sequence database S, and the minimum support threshold min_support.

- **Output** of the algorithm: The complete set of sequential patterns.

- **Subroutine**: PrefixSpan(α, L, S|α).
  
  **Parameters**:
  
  α: sequential pattern,
  
  L: the length of α;
  
  S|α: the α-projected database, if α ≠ < >; otherwise; the sequence database S.

  **Call** PrefixSpan(< >,0,S)

**Method**:

1. Scan S|α once, find the set of frequent items b such that:

   b can be assembled to the last element of α to form a sequential pattern; or <b> can be appended to α to form a sequential pattern.

2. For each frequent item b:

   append it to α to form a sequential pattern α’ and output α’;

   output α’;

3. For each α’:

   construct α’-projected database S|α’ and
call PrefixSpan(α’, L+1, S|α’).
Algorithm 1: FP-tree construction

*Input:* A transaction database DB and a minimum support threshold

*Output:* FP-tree, the frequent-pattern tree of DB.

*Method:* The FP-tree is constructed as follows.

1. Scan the transaction database DB once. Collect F, the set of frequent items, and the support of each frequent item. Sort F in support-descending order as FList, the list of frequent items.
2. Create the root of an FP-tree, T, and label it as “null”. For each transaction Trans in DB do the following:

   - Select the frequent items in Trans and sort them according to the order of FList. Let the sorted frequent-item list in Trans be [p | P], where p is the first element and P is the remaining list. Call insert tree ([p | P], T).

   - The function insert tree([p | P], T) is performed as follows. If T has a child N such that N.item-name = p.item-name, then increment N’s count by 1; else create a new node N, with its count initialized to 1, its parent link linked to T, and its node-link linked to the nodes with the same item-name via the node-link structure. If P is nonempty, call insert tree(P, N)

After constructing the FP-Tree it’s possible to mine it to find the complete set of frequent patterns.
Algorithm 2: FP-Growth

*Input:* A database DB, represented by FP-tree constructed according to Algorithm 1, and a minimum support threshold

*Output:* The complete set of frequent patterns.

*Method:* call FP-growth(FP-tree, null).

Procedure FP-growth(Tree, a)

{
    if Tree contains a single prefix path then // Mining single prefix-path
        FP-tree
        {
            let P be the single prefix-path part of Tree;
            let Q be the multipath part with the top branching node replaced by a null root;
            for each combination (denoted as β) of the nodes in the path P do
                generate pattern β \cup a with support = minimum support of nodes in β;
                let freq pattern set(P) be the set of patterns so generated;
        }
    else
        let Q be Tree;
        for each item ai in Q do
            { // Mining multipath FP-tree

106
generate pattern $\beta = a_i \cup a$ with support = $a_i$ \text{support};

construct $\beta$’s conditional pattern-base and then $\beta$’s conditional FP-tree $\text{Tree} \beta$;

if $\text{Tree} \beta \neq \emptyset$ then

    call FP-growth($\text{Tree} \beta$, $\beta$);

let freq pattern set($Q$) be the set of patterns so generated;

}

return(freq pattern set($P$) $\cup$ freq pattern set($Q$) $\cup$ (freq pattern set($P$) $\times$ freq pattern set($Q$))

}

Source:


2. Wikipedia – Association Rule learning
Sample Questionnaire –for the keywords “database, Database types”

1. What is the component of a DBMS that is responsible for storing, retrieving, and updating data?
   A) data dictionary   B) data management engine   C) database engine   D) query engine

2. The tool which assists in generating input screens is referred to as:
   A) data dictionary   B) forms generator   C) input screen tool   D) report generator

3. The ability to modify the data structure and not have to change the programs using that data is called:
   A) data dictionary   B) data independence   C) data integrity   D) referential integrity

4. Which of the following items is not a DBMS:
   A) Access   B) Acrobat Reader   C) Oracle   D) SQL Server

5. The database design that uses a hierarchical data structure, but incorporates multiple data entry points is called a:
   A) Hierarchical database   B) Network database   C) Object oriented database   D) Relational database

6. Which of the following is an extension of the Relational Database model?
   A) Hierarchical database   B) Multidimensional database   C) Network database   D) Object oriented database

7. What is the main strength of Relational Databases?
A) Ability to handle any type of data   B) Defining objects provides for reuse of data definitions   C) Ease of use   D) Flexibility and efficiency in accessing data

8. Which component of the database management system (DBMS) most affects the performance (speed)?
   A) Data Storage Subsystem   B) Database Engine   C) Query Processor   D) Security Subsystem

9. Data integrity can be improved by which of the following means ____________.
   A) adding indexes   B) entering appropriate terms into the data dictionary   C) incorporating business rules when defining the data   D) using inheritance

10. The role of the query system is to:
    A) present the data in a user friendly format   B) provide data security   C) retrieve and manipulate data   D) support data integrity

11. Which of the following items is not the advantage of a DBMS?
    A) Data Independence   B) Decentralized administration of the data   C) Ease of application development   D) Uniform security, privacy and integrity

12. The term referring to a physical item existing in the real world that you want to track is called:
    A) class   B) entity   C) object   D) table

13. Two different terms are used to describe the characteristics of interest for an entity. They are properties and:
    A) attributes   B) classes   C) entities   D) traits

14. The term used for the functions and procedures that work on class data is:
    A) attributes   B) entity   C) methods   D) objects

15. The database model that utilizes multiple tables interconnected through common attributes to store and manage information is called a:
A) Class Database  B) Matrix Database  C) Network Database  D) Relational Database

16. The process to properly define the database tables to provide flexibility, minimize redundancy, and ensure data integrity is called:

A) class diagram  B) data normalization  C) database design  D) design rationalization
APPENDIX D

Sample e-learning topics of DBMS

Introduction
What is database system, purpose of database system, View of data, relational databases, Database architecture, Transaction management.

Data Models and Architecture
The importance of data models, Basic building blocks, Business rules, The evolution of data models, Degrees of Data abstraction, RAID Architecture

Database Design
ER-Diagram and Unified Modeling Language, Database design and ER Model overview, ER-Model, Constraints, ER-Diagrams, weak entity sets, Codd’s rules, Relational Schemas, Introduction to UML, Relational database model: Logical view of data keys, integrity rules. Relational Database design: features of good relational database design, atomic domain and Normalization (1NF, 2NF, 3NF, BCNF).

Relational Algebra and Calculus

Constraints, Views and SQL
Constraints, types of constraints, Integrity constraints, Views: Introduction to views, data independence, security, updates on views, comparison between tables and views SQL: data definition, aggregate function, Null Values, nested sub queries, Joined relations, Triggers

Transaction management and Concurrency control
Transaction management: ACID properties, serializability and concurrency control, Lock based concurrency control (2PL, Deadlocks), Time stamping methods, optimistic methods, database recovery management.