Mining of Prevalent Ailments in a Health Database Using Fp-Growth Algorithm

Onashoga, S. A. *bookyy2k@yahoo.com*
Dept. of Computer Science,
University of Agriculture, Abeokuta, Nigeria

Sodiya, A. S. *sinaronke@yahoo.co.uk*
Dept. of Computer Science,
University of Agriculture, Abeokuta, Nigeria

Akinwale, A. T. *aatakinwale@yahoo.com*
Dept. of Computer Science,
University of Agriculture, Abeokuta, Nigeria

Falola, O. E. *wunmi3005@yahoo.com*
Dept. of Computer Science,
University of Agriculture, Abeokuta, Nigeria

Abstract

Health databases are characterised by large number of attributes such as personal biological and diagnosis information, health history, prescription, billing information and so on. The increasing need for providing enhanced medical system has necessitated the need for adopting an efficient data mining technique for extracting hidden and useful information from health database. In the past, many data mining algorithms such as Apriori, Eclat, H-Mine have been developed with deficiency in time-space trade off. In this work, an enhanced FP-growth frequent pattern mining algorithm coined FP-Ail is applied to students' health database with a view to provide information about prevalent ailments and suggestions for managing the identified ailments. FP-Ail is tested on a student's health database of a tertiary institution in Nigeria and the results obtained could be used by the management of the health centre for enhanced strategic decision making about health care. FP-Ail also provides the possibility to refine the minimum support threshold interactively, and to see the changes instantly.

Keywords: FP-Ail, Frequent Pattern, Health Database, Knowledge.

1. INTRODUCTION

A good number of domains of life such as the scientific institutions, government agencies and businesses have dedicated a good part of their resources to collecting and storing data, of which only a minute amount of these data will ever be used because, in many cases, the volumes are simply too large to manage, or the data structures themselves are too complicated to be analysed effectively. The primary reason is that the original effect to create a data set is often focused on issues such as storage efficiency, it does not include a plan for how the data will eventually be used and analysed for decision making purposes. Thus, establishing the fact that we are drowning in data but starving for knowledge.

Data mining is defined as the process of extracting trends or patterns from data in a large database and carefully and accurately transforms them into useful and understandable information [2].

Frequent pattern mining has been a constantly addressed factor in the field of data mining as a result of its great and promising applicability at mining association [2], causality [9], sequential patterns [3], just to mention a few.

1.1 Frequent Pattern Mining

Let I be a set of items. A set $X = \{i_1, \ldots, i_k\} \subseteq I$ is called an itemset, or a k-itemset if it contains k items. A transaction over I is a couple T = (tid , I) where tid is the transaction identifier and I is an itemset. A transaction T = (tid , I) is said to support an itemset $X \subseteq I$, if $X \subseteq I$. A transaction database D over I is a set of transactions over I.

A number of algorithms have been proposed for mining frequent patterns in a large database, these include apriori algorithm, pattern growth methods, such as FP-growth [6] and tree projection [7] e.t.c. In a transaction database, a frequent set would be a set of items that co-occur frequently in the database. A pattern growth algorithm, FP-growth, reported to be an order of magnitude faster than apriori algorithm was proposed [6]. In this work, an enhanced FP-growth algorithm is used to mine students' health database by compressing the whole database in a compact manner

This paper is organised as follows: Section 2 has the related works, while section 3 discusses the modified algorithm. Section 4 has the implementation details with the results and section 5 concludes the work with area of further research.

2. RELATED WORKS

[4] describes a scalable, distributed software architecture that is suitable for managing continuous activity data streams generated from body sensor network. The system, when applied to healthcare, helps in taking care of patients' well-being through continuous and intelligent monitoring. The objective was achieved through observation of frequent patterns of the inherent structures of the concerned patients.

CLOTELE, a pattern growth algorithm for mining closed frequent calling patterns of a telecommunication database from a telecommunication provider was proposed in [5]. The knowledge obtained is useful for telecommunication network operators in order to make crucial decisions.

E-CAST used a dynamic threshold and indicated that the cleaning step of the original CAST algorithm may be unnecessary, in which the threshold value was computed at the beginning of each new cluster was introduced [1] . The knowledge gained could be used for the analysis of gene expression data.

[8] presented a systematic approach for expressing and optimizing frequent itemsets queries that involve complex conditions across multiple datasets. This work provided an important step towards building an integrated, powerful and efficient KDDMS which provides support for complex queries on multiple datasets in a KDDMS.(Knowledge Discovery and Data Mining System).

The benefits of performing episodic mining of health data which is a method of compressing transactional set health care episodes, that are standardised medical practise are clearly highlighted in [9]. The benefits include preprocessing data to some temporal principle that is clinically meaningful. It allows for filtering irrelevant attributes that will not be included in data analyses.

3. METHODOLOGY

3.1 Procedure for Mining Frequent Ailments Pattern

Records of different forms such as Patients' billing information, Staff Routine, Patients' health information are kept on daily basis in the health domain which could aid strategic decision making for better health care and profit maximization, if it is well exploited. The procedure used in this research for mining frequent ailments patterns involve the following stages:

1. Data Collection: At this stage, a secondary database of students' health information is extracted from the health management system.

2. Data Cleaning: In order to extract useful frequent ailment pattern, data preprocessing and data cleaning are needed. This stage identifies the most relevant/fundamentally required attributes for mining and also deals with outliers.
3. Pattern Discovery: After the data cleaning stage, the data mining algorithm to discover frequent ailment pattern is designed (section 3.2).
4. Deduction: this involves the comprehensibility of the discovered pattern i.e the conclusion drawn from the discovered knowledge.

StudentID	Gender	Level	Diagnosis	Abode	Dept.	…
A4	M	400	Malaria	Hostel	Maths	…
E3	F	300	Typhoid	Town	ABG	…
F6	F	300	HepatitisA	Town	Stat	…
D9	M	400	HepatitisB	Hostel	Home_ Sc.	…
A4	M	400	Cough	Hostel	Maths	…
…	…	…	…	…	…	…

TABLE 1: An example of selected health database

3.2 Algorithm Design – FP-Ail

This section discusses the FP-Ail algorithm.
Algorithm: Mining frequent ailments pattern, an FP-growth based approach.
Input: Valid (User, Password) Authentication
π (D), σ_{abs} // π (D) is the projection on the database and σ_{abs} is the minimum support.
Output: F(π (D), σ_{abs}) //Frequent ailment patterns.
Methods: The algorithm is given as below:

```
        Login (Username, Password)
        If( Login () is successful )
                then Select π (D)
        Else re-Login()
end if
H= {} // frequent-1 itemsets
TDB={} // Transaction Database
F(π (D), σabs )={}
// create TDB
for all i ϵ π (D) do
    TDB={i, {j}} where  set j corresponds to i = tid
// Prune (Delete infrequent itemsets)
        for distinct j ϵ TDB do
                get Supp (j) // get count of j
            if Supp (j) >= σabs
                H={j, Supp(j)}// get frequent-1 itemsets
             end if
        end for
        for all (tid, X) ϵ π (D) with j ϵ X do
                K= sort(tid(i), Xi ) in Support descending order
        // Depth first recursion
                Compute F[k] //frequent iemsets
        F(π (D), σabs ) = {H U F[k]}
```

FIGURE 1: Fp-Ail Algorithm

3.3 An Illustrative Example for Mining Frequent Ailment patterns

Consider Table 1 for illustration. Suppose the frequent ailments are to be mined based on two attributes: students' ID and diagnosis

Tid	Diagnosis (Itemsets)
A4	m, c, t, d
E3	m, t, u, h, d, t
F6	m, h, d, c, t, m
D9	m, h, c, d

TABLE 2: Transaction Database, TDB

Step 1**-** Compute frequent patterns

The database, from which the transaction database (TDB) is extracted, is allowed to be queried by an authenticated user. This algorithm ensures strict denial to an invalid user as the health domain is always meticulous over privacy issue. However, the user has been saved the stress of the need to write SQL queries by just having to select the required attributes from a populated combo box component of the application. Table 1 is used for illustration, in order to reduce the processing time, the second attribute, diagnosis, is encoded with the first letter of each item in the transaction i.e Malaria= m, Cough= c e.t.c. The TDB generated based on the query is scanned in order to compute the support of each itemset, after which the infrequent itemsets are eliminated based on the given minimum support threshold. (Note: the support is the absolute occurrence of items)

Itemset	Support
C	3
D	4
H	3
M	5
T	4
U	1

TABLE 3: TDB showing item support counts

Note: From the above, u is not frequent and is thus eliminated. (Min. support (σ_{abs}) is set to 3)

Tid	Diagnosis (Itemsets)
A4	m, t, d, c
E3	m, t, t, d, h
F6	m, m, t, d, c, h
D9	m, d, c, h

TABLE 4: TDB sorted in support descending order

Step 2: Use FP-Ail algorithm to mine FP-Ail tree

FP-Ail tree has the information of the whole database, so the algorithm mines this tree and not the database.

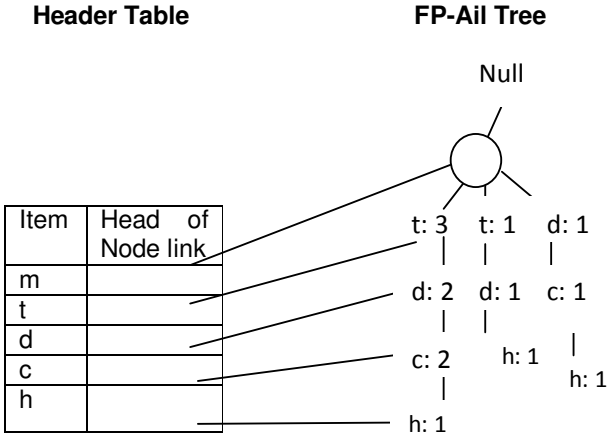
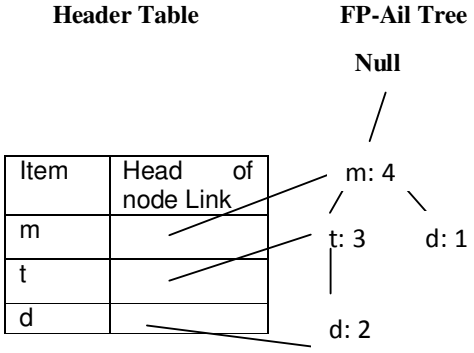


FIGURE 2: The Header Table and FP-Ail Tree

Starting from h, for each frequent-1 itemsets, construct its conditional pattern base. A conditional pattern base for an itemset contains the transactions that end with that itemset.
i). Item h's conditional pattern base is: {m:2, t:1, d:1, c:1},{m:1, t:2, d:1},{m:1, d:1, c:1}.

Note: In this conditional pattern base, c occurs only twice and is thus eliminated.

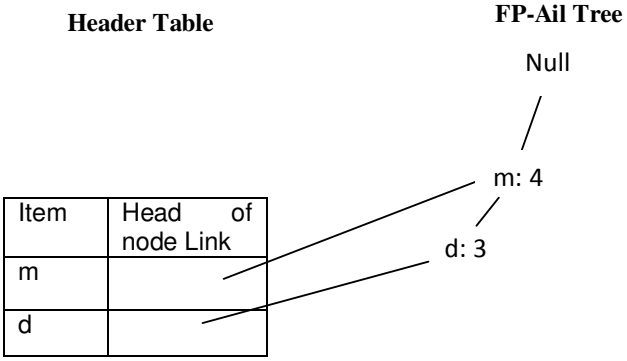
The conditional FP-Ail tree is constructed thus;



The frequent patterns generated for item h is as given below:

{mdh: 3},{dh: 3}, {mh: 3}.

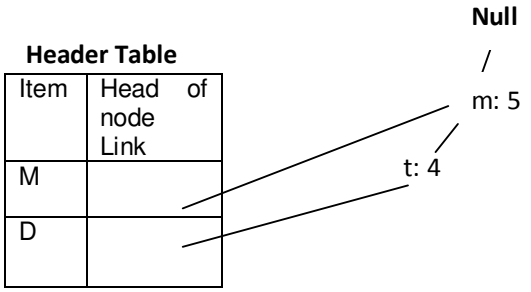
ii). Item c's conditional pattern base is: {m: 1, t: 1, d: 1}, {m: 2, t: 1, d: 1}, {m:1, d:1}.
Note: Items t occurs only twice and is thus eliminated.
The conditional FP-Ail tree is constructed thus:



The generated frequent itemsets is as follows: {mdc: 3}{dc: 3}{mc: 3}

iii). Items d's conditional pattern base is {m: 1, t: 1}, {m: 1, t: 2}, {m: 2, t: 1}, {m: 1}

The conditional FP-Ail tree is constructed thus;



The generated itemsets is as follows: {mtd: 3}, {md: 4}, {td: 3}.

iv). Items t's conditional pattern base is given as: {m: 1}, {m: 1}, {m: 2} and the generated itemsets is {mt: 3}.

Combined with the frequent-1 itemsets generated during the first database scan, we have the following frequent patterns: {mtd: 3}, {md: 4}, {td: 3}, {mt: 3}, {mdc: 3}, {dc: 3}, {mc: 3}, {mdh: 3},{dh: 3}, {mh: 3}.

4. IMPLEMENTATION AND RESULTS

The algorithm is experimented on the student health database of 4 different academic sessions, which is of size 600KB, tested on varying minimum support in order to test the flexibility and adaptability of the algorithm. All experiments were carried out on a 733MHz Pentium III PC, with a 1GB RAM size, 100GB HDD running Microsoft Windows Vista. FP-Ail was implemented in Java using NetBeans IDE 6.1 version.

The system designed is so flexible in that it allows different attributes to be selected based on the operational environment and the expected knowledge to be discovered. In particular, during this implementation, three different database is selected from the TDB for mining. These databases are T60I15D100K, T4I15D100K and T10I15D100K where T represents students' ID: 60, different levels of students: 4 and different students' abodes: 10 respectively and I is the number of diagnosis and D represents the number of records in the data base. The student ID is mined against diagnosis.

4.1 Types of Knowledge Discovered

The knowledge to be acquired from health databases can not be over-emphasized. However, in this discourse, from the sequence of diagnosis pattern, the management could discover the most effective therapy, know the most likely ailment a particular patient could have from the health record or determine the most affected group of patients with a particular ailment in order to make strategized and optimal decisions. Table 5 shows examples of patterns extracted from respective databases with their interpretations and several decision that could be taken by the authorities concerned.

	Patterns	Interpretation	Decision
T60I15D100K	{H8:cough, Tuberculosis: 8}	Student with ID H8 is noticed to have been diagnosed with the ailments in that order on 6 occasions.	Student's health should be monitored and may need to be sent home for treatment in order to avoid spread of the diseases.
T4I15D100K	{200L: diarrhoea, malaria: 20}	200 level students were been attacked severally with the identified ailments	The pharmaceutical unit should be stocked with drugs and measures should be taken to reduce or combat the attack.
T10I15D100K	{Asero: diarrhoea, malaria: 18}	Several students living at Asero (an abode outside the campus at Abeokuta, Ogun State, Nigeria) were discovered to have been frequently diagnosed with the listed ailments.	The university management could call the attention of the government in carrying out a health inspection of the area and provide necessary measures.

TABLE 5: Interpretation of the patterns extracted

4.1.1 Prevalent Ailments

We went further to mine the prevalent ailments among students by considering each session starting from 2003/2004. This window (Figure 3) displays the result of the successful query, for example:

SELECT *STUDENTID, DIAGNOSIS*
FROM *TDB*
WHERE *LEVEL = 100*

on the database. The patterns generated are displayed in the lower right end component of the windows displayed. This result would go a long way in assisting the health management make crucial decisions.

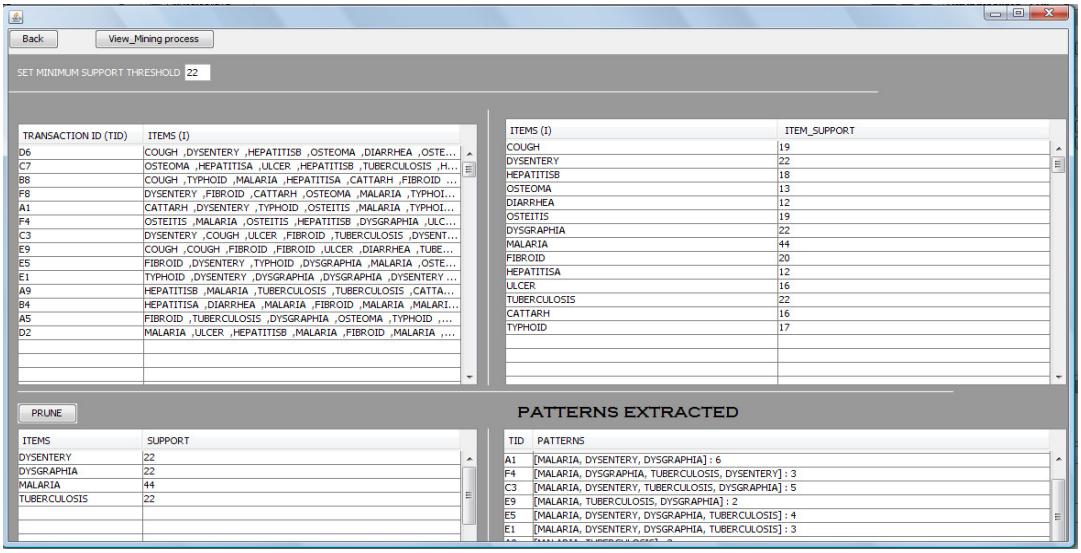


FIGURE 3: Mining View with Patterns generated.

The result as depicted in Table 6 using diagnosis against the year showed that {Malaria, Headache} was rampant among the students in 2003/2004 and 2005/2006. This clearly

shows that the common ailments as reported in the health database is this pattern. In this regards, the management should try and design a measure of reducing stress on the parts of students which could be the cause of headache and find a way of reducing the attack of malaria either by always fumigating the environment at the end of each semester or provide some mosquito repellent tools.

	SUPPORT		
YEAR	*10*	*15*	*20*
2003/2004	{Malaria, Headache, cold, catarrh, diarrhoea, dysentery, hepatitis}	{Malaria, Headache, cold, diarrhoea}	{Malaria, Headache}
2004/2005	{Malaria, diarrhoea, catarrh}	{Malaria, diarrhoea, catarrh}	{Malaria}
2005/2006	{Malaria, Headache, tuberculosis, dysentery, catarrh}	{Malaria, Headache, tuberculosis, catarrh}	{Malaria, Headache, catarrh}
2006/2007	{cold, pains}	{cold, pains}	{ pains}

TABLE 6: Patterns on prevalent ailments with support values

The choice of the minimum support value is critical in many applications, if it is too high, the FP-Ail tree is empty, if it is too low, the number of items in the FP-Ail tree is too high. Therefore FP-Ail provides the possibility to refine the minimum support threshold interactively, and to see the changes instantly. Figure 4 depicts each of these databases in comparison with the changing of support. The size of the patterns generated is inversely proportional to the minimum support.

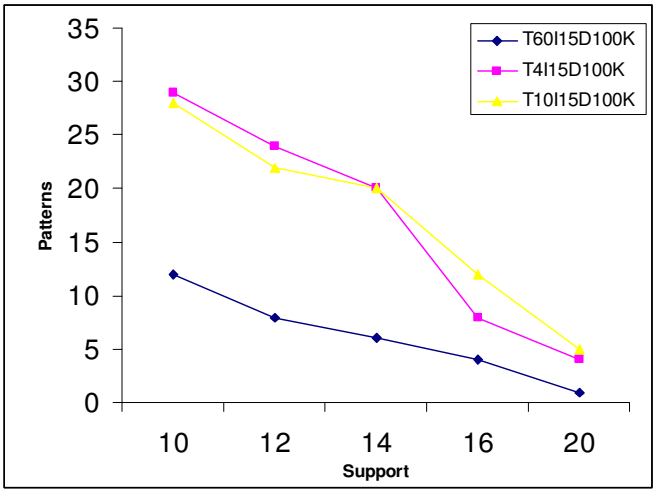


FIGURE 4: Effects of changing support threshold

5. CONCLUSION

In this paper, we have highlighted the advantages of FP-growth as a frequent pattern mining algorithm. It is thus integrated into an undergraduate students' health database coined FP-Ail algorithm for extracting hidden knowledge that could aid strategic decision making in the health unit of any organization. The tool designed is flexible and integrates security based on privacy issues of the health domain. The algorithm is tested and several knowledge is discovered.

ACKNOWLEDGEMENT

Our appreciation goes to the Director of the University Students' Health Centre for giving us assess to the database.

REFERENCES

[1] B. Abdelghani, P. David, C. Yidong, G. Abdel (2004). "E-CAST: A Data Mining Algorithm For Gene Expression Data". *BIOKDD02: Workshop on Data Mining in Bioinformatics with SIGKDD02 Conference.*

[2] R. Agrawal and R. Srikant (1994). "Fast Algorithms for mining association rules", *Proceedings 20th International Conference on Very Large Data Bases, pages 487- 499. Morgan Kaufmann.*

[3] R. Agrawal and R. Srikant (1995). "Mining sequential patterns". *ICDE'95.*

[4] R. Ali, M. ElHelw, L. Atallah, B. Lo, Y. Guang-Zhong. (2008). "Pattern mining for routine behaviour discovery in pervasive healthcare environments". *International Conference on Technology and Applications in Biomedicine, 2008. ITAB 2008.*

[5] S. A. Ibrahim, O. Folorunso, O. B. Ajayi (2005). "Knowledge Discovery of Closed Frequent Calling Patterns in a Telecommunication Database". *Proceedings of the 2005 Information Science and IT Education Joint Conference. Flagstaff, Arizona, USA. June 16-19.*

[6] J. Han, J. Pei, and Y. Yin. (2000). "Mining Frequent Patterns without Candidate Generation". *Proc. 2000 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD'00)*

[7] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, & M. C. Hsu (2001). "Prefix-Span: Mining sequential patterns efficiently by prefix-projected pattern growth". *Proceedings 2001International Conference Data Engineering (ICDE'01).*

[8] J. Ruoming, A. Gagan (2004). "A Systematic Approach for Optimizing Complex Mining Tasks on Multiple Databases". *Department of Computer Science and Engineering, Ohio State University, Columbus OH 43210.*

[9] T. Semenova (2003). "Episode-Based Conceptual Mining of Large Health Collections" *Lecture notes in Computer Science, Vol. 2813/2003, Publisher: Springer Berlin / Heidelberg.*