2012 International Conference on Information and Knowledge Management (ICIKM 2012)
IPCSIT vol.45 (2012) © (2012) IACSIT Press, Singapore

Information Extraction using Automated Wrapper Builder using Approximate Occurrence Identification Algorithm for Health Systems

Samir K Amin[1], Khairuddin Bin Omar[2] and Dinesh Kumar Saini[3]

Abstract. In the paper we are building the wrapper automation using approximate occurrences identification algorithm which is very much needed in the health systems. The input to IIEHC is a set of one-record or multiple-record text documents for patients' medical records. The enclosing operation can be applied to one (or several if necessary) of the training documents. The research describes how the records in other training documents are recognized.

Key Words: Information Extraction, Wrappers, Altova

1. Introduction

The user can specify the number of records in the enclosed block. Let k be the number of records in the enclosed block. For k = 1 (default), the encoded token string of the enclosed block is saved as the primitive record pattern, P. For k > 1, the following method is to discover the patterns that occur k times in the enclosed block and then generalizes the segments between two adjacent occurrences by multiple string alignment. The record pattern is then expressed as a signature representation R which contains alternatives. Note that the method described with k=1 finds approximate occurrences of a string P in a training page T. Here, we have a regular expression R instead of a simple string P. Therefore, the research extends R to the longest expression P (without spaces and alternatives) and applies the same technique of k=1 to identify approximate occurrences of P. For example, for a regular expression R = dtb [a |-] t [b |-] [t |-], the longest expression is P = dtbatbt.

2. System Building

To discover other records in the input pages for further training, exact pattern matching is of no help in providing new information. Instead, the research expects the discovery of a substring that is similar to P so that we can generalize these substrings for extending the pattern. The research develops a precise definition of similarity between two strings S1 and S2 as follows [3].

Definition 2.1 *Let Σ be the alphabet used for strings S1 and S2, and let Σ' be Σ with the added character"-"denoting a space. Then, for any two character x , y in Σ' , the function match(x,y) denotes the value obtained by aligning character x against character y.*

Definition 2.2 *An (global) alignment of two strings S_1 and S_2 is obtained by inserting chosen spaces either into or at the ends of S_1 and S_2, such that the resulting strings S'_1 and S'_2 are of equal length. The value of such an alignment is defined as,*

$$\sum_{i=1}^{l} match(S'1[i], S'2[i])$$

where l denote the (equal) length of the two strings S'_1 and S'_2 in the alignment.

Definition 2.3 *The similarity score, $SC(S_1,S_2)$, of two strings S_1 and S_2 is the optimal value of all (global) alignments between S_1 and S_2. We also define similarity ratio as the similarity score divided by s.min{$|S_1|,|S_2|$}, the maximum value by matching S_1 and S_2.*

$$V(i,0) = i*d; \quad (1)$$

and

$$V(0, j)=j*d; \quad (2)$$

where $V(i,j)$ be the value of the optimal alignment of $S1[1..i]$ *and* $S2[1..j]$.

For i and j, the general recurrence is $V(i-1,j-1)+match(S_1[i]S_2[j])$;

$$V(i,j) = max \begin{cases} V(i-1, j)+d; \\ V(i,j-1)+d; \end{cases} \quad (3)$$

Definition 2.4 *Given a parameter ($0 < \theta < 1$), a substring T' of T is said to be an approximate occurrence of P if, and only if, the similarity ratio of P and T' is at least θ, or the similarity score is at least $\delta=\theta$. s.|P|, where s.|P| denotes the largest value matching pattern P.*

Theorem 2.5 *There is an approximate occurrence of P in T ending at position j of T if and only if $V(n,j) \geq \delta$, where n is the length of P. Moreover, T[k,j] is an approximate occurrence of P in T if, and only if, $V(n,j) \geq \delta$ and there is a path of backpointers from cell (n , j) to cell (0,k).*

Definition 2.6 *The sum of pairs (SP) score of a multiple alignment M is the sum of the similarity scores of pairwise global alignments induced by M, where the induced pairwise alignment is defined as follows.*

Definition 2.7 *Given a multiple alignment M, the induced pairwise alignment of two strings S_i and S_j obtained from M by retaining S_i and S_j and removing any two opposing spaces in these two rows.*

3. Wrapper Key Constructor

The output text files containing the extracted information present to users for their opinions through the interface, once the user agrees to the results, the system will store the resulting wrapper into a pool after constructing a key to it by the WKeys Constructor procedure, this key is the sequences of medical abbreviations according to the tree in Figure 6 for the information extracted by that wrapper. For example the key: <01><04><05> will enable a wrapper to extract information regarding AGE, SBP, and DBP fields from medical records. The Key to String Encoding procedure will take this key as input and produce binary bit string according to the position of the fields as shown in Figure 1.

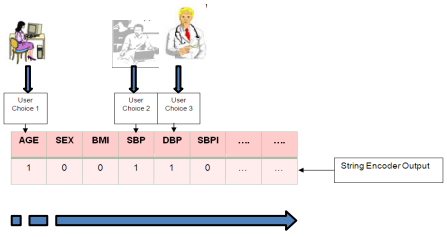

Figure 1 Key to String Encoder Output

This new string will be added as a new entry in a table called the WKeys Table by a procedure called the WKeys Table Constructor. Once added, the table will be passed as an input parameter to another procedure called the Patterns Table Constructor which will build a two dimensional symmetric table as in Figure 2

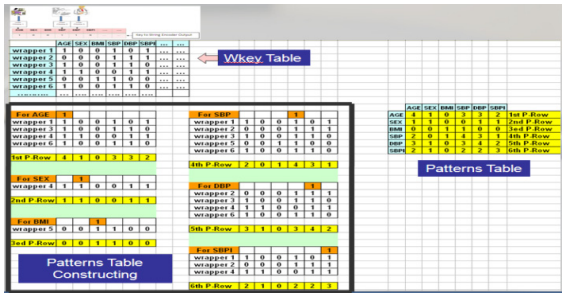

Figure 2 Patterns Table Construction Process

The new wrapper will be written to the wrapper pool, while the amended Patterns Table will be written as new meta-knowledge as shown in Figure 3.

3.1. System Framework (Wrapper Retrieval Part)

The Wrapper Retrieval Component of the system includes two main components; a graphical user interface called pattern viewer, which shows repetitive patterns discovered and represents a communication environment with the users; and the wrapper key retrieval component which accepts an input queries and contains the core techniques of pattern mining which is implemented in the wrapper key retrieval.

3.2. Wrapper Key Retrieval

This part includes; Consistency Query Handling; Key to String Encoder; Pattern Allocator; Pattern validator; and a Wrapper Presenter to user. (Figure 3)

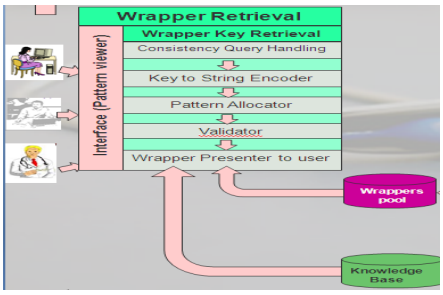

Figure 3Wrapper Retrieval part

Wrappers are presented to users by the wrapper presenter using the graphical user interface enabling users to view and edit those wrappers. Once the user selects a target wrapper conforming to his/her information desires, the extractor module can use the wrappers to extract information from other pages having similar structures with the input page. Figure 3 presents a flowchart of the wrapper retrieval component. The flowchart presents an overview of the wrapper key mining process; more details will follow as to how a target key pattern for wrappers is generated. When users submit queries to the system through Consistency Query Handling procedures, the Key to String Encoder will receive the query and translate it into a string of abstract representations referred to here as WKeys. Each WKey is represented by a binary code of fixed length (See Figure 1). The Pattern Allocator then uses the PatternsTable to discover wanted patterns. After users agree, the wanted patterns are forwarded to the validator, which filters out undesired fields and produces the wanted wrapper file name. Finally, the wrapper presenter retrieves the wrapper file to present it to the user for final confirmation. The system uses the knowledge base as meta-knowledge concerning the stored wrappers in the wrappers pool. In Figure 4, there is a representation of a sample Patterns Table. It is that this table is a square symmetric table, Its diagonal numbers represent the number of wrappers in the pool that are capable of extracting information representing the combination of the row and column fields.

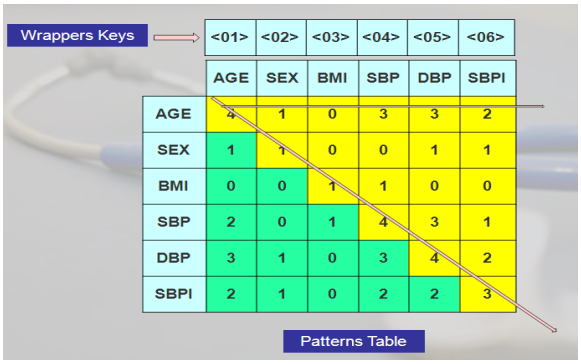

Figure 4 Sample Patterns Table

For example, we have in the wrapper pool 4 wrappers dealing with the AGE field, with the SEX field and so on. We have no wrappers dealing with BMI and SEX fields together (because we have 0 in the intercept of these two fields). Once the candidate wrappers are discovered, the user may select from these

candidates his target wrapper that contains desired information. The extractor receives a target wrapper file key as input and applies that wrapper to the documents to extract the desired information.

3.3. User Interface (Pattern Viewer)

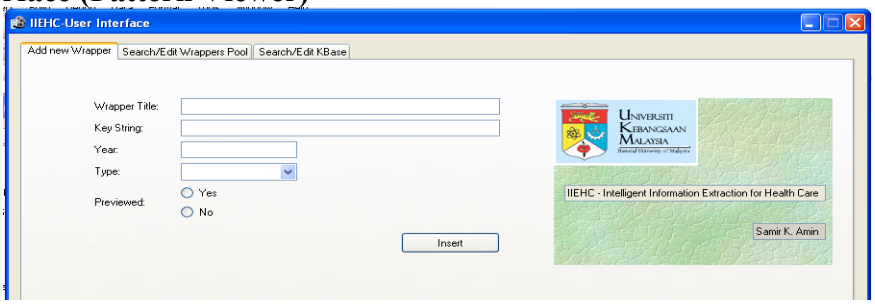

Figure 5 User Interface (Pattern Viewer)

Tag tokens can be classified in many ways. The user can choose a classification depending on the desired level of information to be extracted. For example, in the case of medical information, it is sufficient to associate numbers for the medical abbreviations as in Figure 5 above.

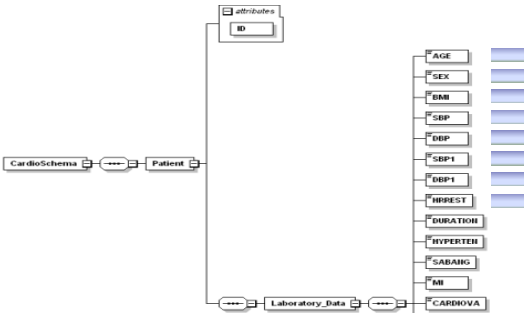

Figure 6 Numbering the medical abbreviation using the Tree

The many different tag classifications allow different wrappers to be generated. With these different abstraction mechanisms, different patterns can be produced. For example, skipping all text-level tags will result in higher abstraction from the input document than all included tags. In addition, different patterns can be discovered and extracted when different encoding schemes are translated.

3.4. Meta-Knowledge Report

Figure 7 illustrates the patterns stored in the knowledge base as meta-knowledge concerning wrappers in the system and the fields each wrapper can extract.

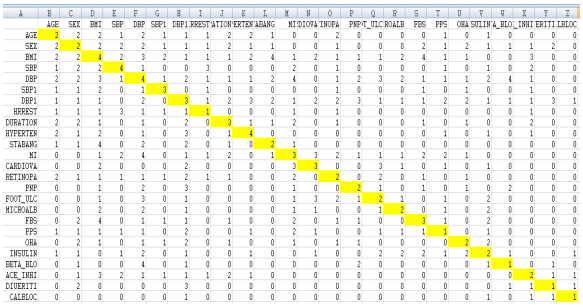

Figure 7 Knowledge Base table

Figure 8 shows the Altova representation of each wrapper in terms of the medical fields it can extract while Figure 9 shows the output report from these wrappers.

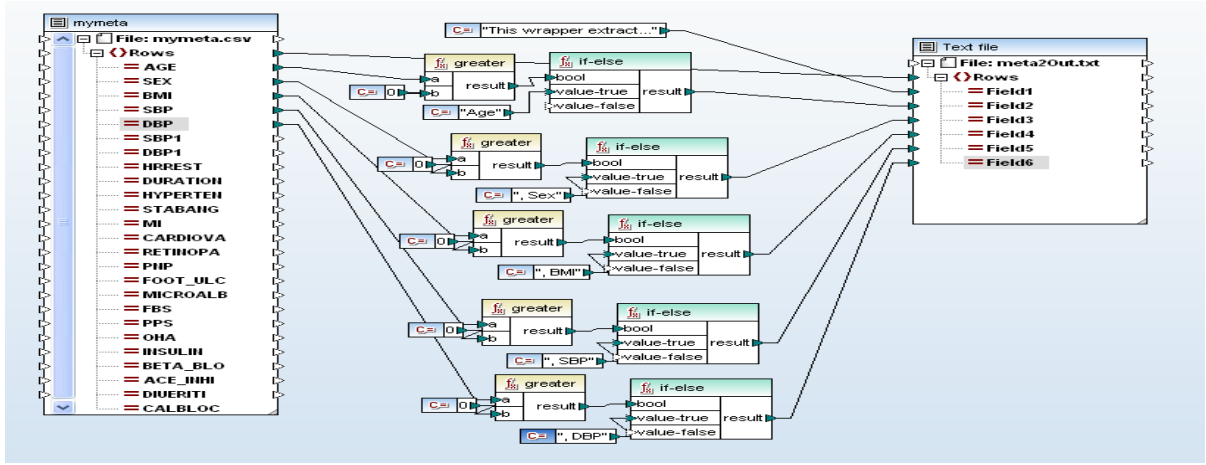

Figure 8Altova wrappers for producing a report.

```
1   This wrapper extract's: AGE , SEX , BMI , SBP , DBP , SBPI , DBPI , HRREST , DURATION , HYPERTEN , STABANG   , RETINOPA     , PPS  , INSULIN
2   This wrapper extract's: AGE , SEX , BMI , SBP , DBP , SBPI , DBPI , HRREST , DURATION , HYPERTEN , STABANG   , RETINOPA    , FBS , PPS , OHA
    , INSULIN , BETA-BLO , ACE-INHI , DIURITI
3   This wrapper extract's: AGE , SEX , BMI , SBP , DBP , SBPI , DBPI , HRREST , DURATION , HYPERTEN , STABANG , MI , CARDIOVA , RETINOPA , PNP ,
    FOOT-ULC , MICROALB , FBS , PPS , OHA   , ACE-INHI
4   This wrapper extract's: AGE , SEX , BMI , SBP , DBP   , HRREST    , MI  , RETINOPA     , PPS  , INSULIN  , ACE-INHI
5   This wrapper extract's: AGE , SEX , BMI , SBP , DBP , SBPI , DBPI , HRREST , DURATION , HYPERTEN , STABANG , MI  , RETINOPA , PNP , FOOT-ULC
    , MICROALB , FBS , PPS , OHA , INSULIN , BETA-BLO , ACE-INHI
6   This wrapper extract's: AGE , SEX , BMI  , DBP , SBPI  , HRREST      , RETINOPA    , FBS  , OHA   , ACE-INHI
7   This wrapper extract's: AGE , SEX , BMI  , DBP  , DBPI , HRREST , DURATION , HYPERTEN , STABANG , MI , CARDIOVA , RETINOPA , PNP , FOOT-ULC ,
    MICROALB , FBS , PPS , OHA , INSULIN , BETA-BLO , ACE-INHI , DIURITI
8   This wrapper extract's: AGE , SEX , BMI , SBP , DBP , SBPI , DBPI , HRREST    , MI  , RETINOPA         , ACE-INHI
9   This wrapper extract's: AGE , SEX , BMI  , DBP  , DBPI  , DURATION , HYPERTEN , STABANG , MI  , RETINOPA    , FBS  , OHA   , ACE-INHI
10  This wrapper extract's: AGE , SEX , BMI  , DBP  , DBPI  , DURATION , HYPERTEN         , PPS  , INSULIN  , ACE-INHI
11  This wrapper extract's: AGE , SEX , BMI  , DBP  , DBPI  , DURATION  , STABANG , MI
12  This wrapper extract's:    , BMI , SBP , DBP  , DBPI , HRREST , DURATION  , STABANG , MI , CARDIOVA , RETINOPA , PNP , FOOT-ULC , MICROALB ,
    FBS , PPS , OHA
13  This wrapper extract's:    , BMI    , DBPI     , MI , CARDIOVA   , FOOT-ULC , MICROALB  , PPS  , INSULIN
14  This wrapper extract's: AGE , SEX , BMI , SBP , DBP , SBPI , DBPI , HRREST , DURATION   , MI  , RETINOPA  , FOOT-ULC  , FBS  , OHA
15  This wrapper extract's:   , BMI  , DBP  , DBPI     , MI   , PNP , FOOT-ULC  , FBS  , OHA  , BETA-BLO
16  This wrapper extract's:   , BMI  , DBP  , DBPI     , MI , CARDIOVA , RETINOPA , PNP , FOOT-ULC , MICROALB  , PPS  , INSULIN
17  This wrapper extract's:   , BMI  , DBP  , DBPI     , MI , CARDIOVA   , FOOT-ULC , MICROALB  , PPS  , INSULIN
18  This wrapper extract's:  , SEX , BMI  , DBP , SBPI , DBPI  , DURATION   , MI  , RETINOPA , PNP   , FBS , PPS  , INSULIN
19  This wrapper extract's: AGE , SEX , BMI , SBP , DBP  , DBPI   , HYPERTEN  , MI , CARDIOVA   , FOOT-ULC , MICROALB , FBS , PPS  , INSULIN
20  This wrapper extract's:  , SEX , BMI  , DBP , SBPI , DBPI  , DURATION   , MI  , RETINOPA , PNP     , OHA , INSULIN
21  This wrapper extract's: AGE , SEX  , SBP , DBP  , DBPI   , HYPERTEN   , CARDIOVA   , FOOT-ULC , MICROALB , FBS , PPS , OHA , INSULIN ,
    BETA-BLO
22  This wrapper extract's:  , SEX   , DBP  , DBPI        , PNP      , INSULIN , BETA-BLO  , DIURITI
23  This wrapper extract's:  , SEX , BMI , SBP , DBP , SBPI , DBPI , HRREST , DURATION , HYPERTEN            , ACE-INHI , DIURITI
24  This wrapper extract's:  , SEX     , DBPI               , BETA-BLO , ACE-INHI , DIURITI
25  This wrapper extract's:       , DBPI              , INSULIN  , ACE-INHI , DIURITI
```

Figure 9 the output report

4. References:

[1] Baumgartner, R.,Flesca, S., &Gottlob, G. (2001). *Visual Web information extraction with Lixto*. In *Proceedings of the 27th International Conference on Very Large Data Bases* (pp. 119–128). Roma, Italy.

[2] Grieser, G.,Jantke, K.P., &Lange, S. (2002). *Consistency queries in information extraction*. In *Proceedings of the 13th International Conference on Algorithmic Learning Theory* (pp. 173–187). LNAI 2533. Lübeck, Germany.

[3] Grieser, G., &Lange, S. (2001). *Learning approaches to wrapper induction*. In *Proceedings of the 14th International FLAIRS Conference* (pp. 249–253). Key West, FL.

[4] Jantke, K.P., &Beick, H.R. (1981). *Combining postulates of naturalness in inductive inference*. *Elektronische Informationsverarbeitung und Kybernetik*, 17 (8/9), 465–484.

[5] Knoblock, C.,Lerman, K, Minton, S., &Muslea, I. (2000). *Accurately and reliably extracting data from the Web*. *Data Engineering Bulletin*, 23(4), 33–41.

[6] Kushmerick, N. (2000). *Wrapper induction: Efficiency and expressiveness*. *Artificial Intelligence*, 118(1-2), 15–68.

[7] Lange, S.,Grieser, G., &Jantke, K.P. (2003). *Advanced elementary formal systems*. *Theoretical Computer Science*, 298, 51–70.

[8] Lerman, K.,Minton, S., &Knoblock, C. (2003). *Wrapper maintenance: A machine learning approach*. *Journal of Artificial Intelligence Research*, 18, 149–181.

[9] W Liu, X Meng (2010) "A vision-based approach for deep web data extraction" Knowledge and Data Engineering, ieeexplore.ieee.org