An Automatic Health Record Ranking by Relative Importance to Current Situation

Nyamsuren Ganbold, Sangwook Kim

School of Electrical Engineering and Computer Science, Kyungpook National University, Daegu, South Korea
suren@media.knu.ac.kr, kimsw@knu.ac.kr

Abstract. This study explores methods to identify the relationship between human activity and health condition and number of possibilities those could be provided by this relationship. This research proposes human activity pattern detection method along with model of degree of importance of health record. This research utilizes the person's life pattern analysis based on activities of daily living. Proposed anomalous activity detection and representation method analyzes one's life pattern by learning patterns through customized decision tree.

Keywords. Ranking, Degree of Importance, Context-Awareness, Importance Mining, Personal Health Record, Decision Tree, Health Data Retrieval

1 Introduction

People leave behind tracks of their lives as they live their lives. One's life history and living pattern can be very helpful in maintaining one's health. Currently people track their health data in format so-called health records. There are many types of health records are available today, ones those managed by Health Care Providers (HCP) and Personal Health Records (PHR), which managed by patients themselves. In this research this study tries to enhance the human technology interaction, accessibility, and usability of those records in case of emergent needs, as fast information retrieval in health care is extremely crucial when it comes to saving person's life. Only highly important and relevant information from their PHR is required for saving patient's life, in case of medical emergency.

Measuring the importance of a record is not an easy task. In real life, if a document has constant access to it, then it considered to be relatively important. Reason is that content in that data is important, so people seeks for it whenever its needed. The more seek counts the record has, the more it might be important. Given this assumption, we can reason that if certain data is important to a user, then user tends to use or access that data a lot more than less important ones. Another Importance of data is not consistent. It is subject to current situation and user's current needs. From this we got our second assumption, that if the data is access frequently in certain situations, we can say that, the data is relatively more important for that specific situation. Then what's left is how to calculate that relative importance of a data.

In order to estimate relative degree of importance of patient's health record, it needs to have system that can handle patient's access history on his health records and analysis on his living pattern. Analysis on patient's living pattern can tell significant insight on patient's health pattern. If patient's living pattern is analyzed successfully, then we can estimate the given documents degree of importance for given pattern.

Thus, goal of this research is to design the system that can rank health records for the given situation by degree of importance, rapidly. Relative of importance of a record is calculated using two factors of patient's health record, Activeness, and Relevancy along with anomaly analysis in patient's life pattern. This research describes the proposing ubiRank, a method for rating health records intuitively using careful life pattern analysis. Also it tries to design context aware personal health record system, the testbed environment for ubiRank.

2 Related Works

For our best information to date, there has been no same research has been done on relative importance ranking of health data. However, many researchers had given address to problems of ranking of web pages and general documents. Most noteworthy from this field are a method of web page ranking, PageRank [2] and topic sensitive ranking [6].

PageRank [2] uses only link structures in the web pages to identify the importance of a webpage. The basic idea is that important page A is directing to page B, implies that page B is also important. If a page is cited by another important page, then the page is also important. This study used recursive algorithm on inbound and outbound links on web pages to rank pages. Basic principle of PageRank is shown in Figure 2.1. Topic sensitive ranking [6] focused on topics, and it made efforts to improve original PageRank [2] algorithm. H. Small studied and analyzed the measure of relationship between two documents [1]. In the research [5] on analysis of degree of importance by M. Murata et al studied importance of data using machine learning techniques. In Table 2.1 comparison of existing document ranking algorithms developed in the past.

Notable research on similarity between two objects conducted by A. Tversky [4]. This research gives important insights on similarity analysis of objects from psychological point of view. Another interesting and noteworthy research [3] explored the similarity measurements between two documents and proposed SimRank. SimRank studied some important aspects of similarities between two objects.

3 Relative Importance Estimation of Records in Personal Health Record

In this section, it describes the proposed models and methods in detail. First part presents human life pattern presentation model introduces several important definitions defined in this study. Then, it presents about the design of degree of importance model for ranking of health records. Importance of a record is estimated from usage log data on patient's health record and anomalous patterns in patient's life. For this purpose, we designed the method of anomaly detection in patient's living pattern and described it in detail in the following chapter.

3.1 Overview of Mobile Personal Health Record in Context — Aware Environment

Health records in this thesis refers to any health record, or event that documented inside patient's PHR, that tracks any change in health condition as in Figure 3.2. For example, this includes doctor's note on patient's visit or patient's dairy entry that records health change or diagnosis made recently. In Figure 3.2, between the each health records (R), there are detected anomalous activities.

This context aware PHR is armored with tool to track patient's activities of daily living activities. These records, Observations of Daily Living [20] is used to analyze patient's living pattern and looks for any existing connection between health condition and life style of patient. The word patient in this dissertation is not limited to only people who are suffering from chronic conditions, and it also includes regular people who try to stay healthy.

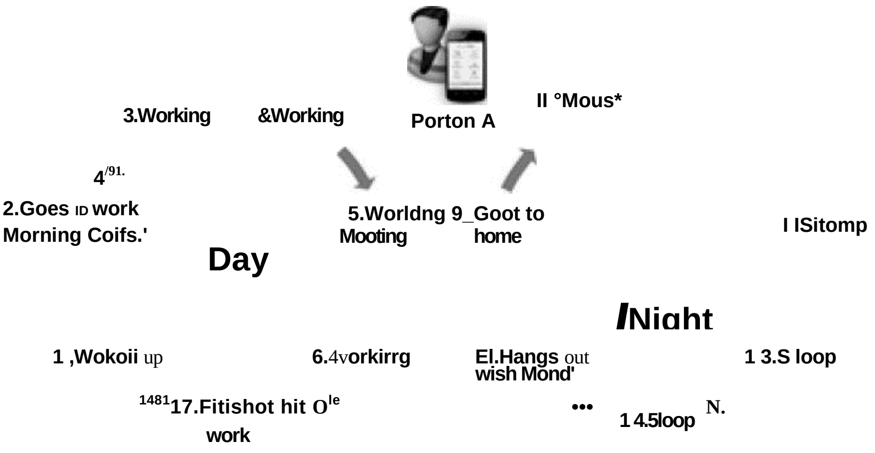


Fig. 1. Example of human activity

pattern 3.2 Anomalous activities in Life: Modeling Life **Pattern**

One's daily living activity pattern can be represented by series of recurring events and activities. Those recurring activities tend to happen in pattern and this research tries to identify least existing patterns in them. Consequently, we are focusing on the detection of anomalous patterns in person's daily living.

Designing life pattern. In order to detect rare or outlier activities in life pattern, we need to know the person's pattern of ADL. Patterns of human life can change periodically based on events happened in their lives. For specific periods in time, they regulates similarly and changes little over time.

Definitions of life pattern and representation of ADL in this thesis is described below.

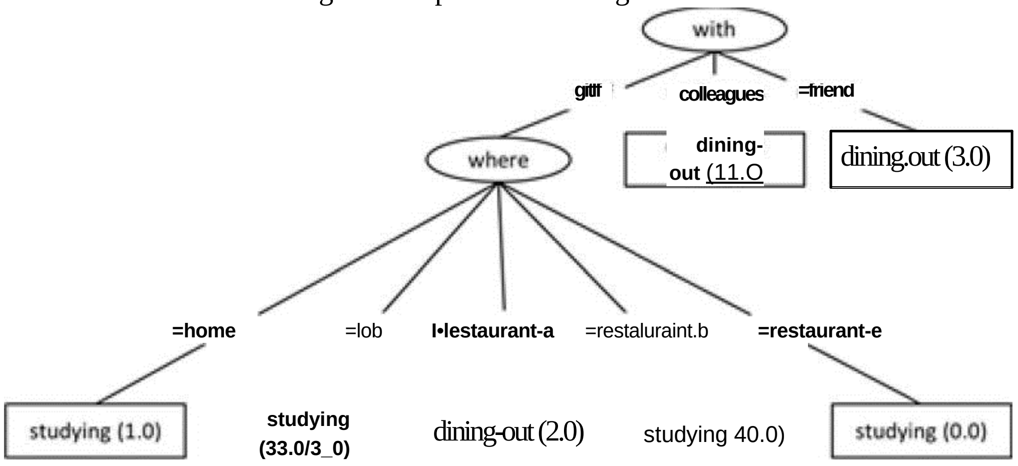
Definition 1. Person's Mean Living Pattern is defined as sequences of One's Activities of Daily Living (ADL). *MLPS* being mean life pattern sequence. MLPS consists of sequences of activity *E* total of *n*.

$$MLPS = \quad E_2, E_{n-1}, E_n) \tag{1}$$

Definition 2. Activities in Person's Mean Living Pattern are defined by Activity Description String (ADS). ADS is encoded sets of activity category codes.

$= \{L_1, \mathbf{L_2}, \mathbf{L_3}, \mathbf{L_4}\}$. Category codes of activities are unique for each individual and represented by capitalized character. There are 4 types of categories in this model each answers to questions, "what?", "with?", "where" and "when" respectively. Precedence of letters are not interchangeable.

Fig. 2. Example decision tree generated



ple decision tree

Definition 3. Anomalous Activities are the activities those lay out of person's mean life pattern.

$$AE, = tAE_1,AE_2, \qquad AE \sqsubseteq E \tag{2}$$

where n is total number of anomalous activities. ❑

Definition 4. Let c_i be situational context collected by context collector for activityE_1. Then contexts for $E_i$defined as, $C(E) = \{c_1, c_2, \ldots, c_{ri}\}$ where nis total number context in context space. o

Definition 5. Mean living pattern is learned by customized decision tree structure and estimates most likely activity for given situational contexts. Activity *E* is target for learning living pattern and dependable variable for given situational contexts c_1.

$$(c, E\ (L_i)) = \qquad c_2, \ , c_k, E\ (L_i)) \qquad (^3)$$

Figure 2 is example decision tree to design our mean living pattern tree defined. The representation of one's mean living pattern is chosen as customized decision tree style because, one's living pattern is highly variable, even though, one has specific living pattern over some periods of their lives. That makes decision tree representation more appropriate

Mean Life Pattern Estimation: Tree of MLP. A typical man's living pattern changes over certain periods of his life time. To adapt above mentioned mean living pattern variation problem, we redesigned classic decision tree algorithm known as C4.5 [20] first introduced by Quinlan as ID3 [19]. Person's patterns for activities of daily living can be learned by decision tree, if there exist sufficient amount of history data for activities performed by person. With the available knowledge of person's ADL dataset defined in format as Definition 5.

Let $D =$ ..., E_n}be dataset containing each activities done by user U. Then each E_i is defined as below,

$$E_i = \qquad c_2, \ , c_k, E, (L_i)), \text{ where, } c_k, L_i \text{ E } C. \qquad (4)$$

With this given dataset D, we can now generate, decision tree for mean life pattern, DT_{mip}. The tree is generated by Algorithm 3.1. Information Gain is used to see if the given context is important in doing specific activity. Information gain is defined as in equation (6) as in ID3, C4.5 algorithms [19], where gain is based on the concept of entropy. In equation (6), P_i is probability frequency of item i's appearance in D.

$$Gain(X) = \qquad P_i log_2 P_i \qquad (^5)$$

Definition 6. Expected MLP can be expressed by sequences of activities, described in ADS, after each event has happened. As time progresses, each activity E is recorded and creates set of actual activities *E actual,* and set of activities predicted by MLP tree, Ep_{redict} .

$$Eactual = \qquad .\bullet\bullet\ , E_{a,n}\} \qquad (6)$$

$$Epredict = (E_{m1}, \ , E_{m\ n}) \qquad (^7)$$

where n is total number of activities done. This set of activities, later will be used in detecting anomalous events in them. o

Anomalous Activities in life. Anomalous activities in life are the ones those are outlier in mean living pattern. Anomalous activities are more likely to lead to future problems and unexpected events. Anomalous events between health records are stored as sequences of events. Events in Anomalous event sequence are included in the set if its value exceeds the given threshold

Anomalous activities are kept as sequences of activity description strings as shown below.

$$E_{predict} = \{BDHC, LDCA, JECD, HEOD, AGHI, BAHC\}$$

$$F \quad \{BDHC, LDCA, JECD, HEOD, AGHI, BAHA\}$$

3.3 ubiRank — Degree of Importance Model

Relative importance of a record is highly dependent on user's usage history on that document along with situational data attached to it. Collecting those entire useful context data and usage log alone is such complex task, even if we suppose that we had all the data we needed, reasoning the degree of importance by factors is much more complicated.

Various factors can be used to determine degree of importance of given document in given situation. Certain record can be considered important in given situation if it was heavily used in similar situations in the past. The main intuition behind ubiRank is that specific record's degree of importance can be calculated, if that record was used in continues and active manner, then degree of importance of that data can be calculated. Degree of Importance of data depends on how actively used the data itself by its owner and for the current situational group. If interaction on the data is increased during specific context group, then it can be said that the data is important in certain situational group.

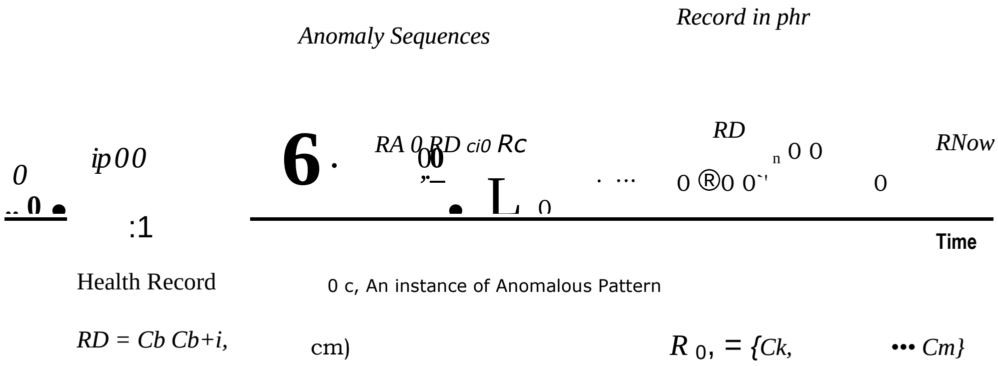


Fig. 3.Two records in timeline

Degree of Importance. Naturally, importance for information is different for different people in various situations. In other words, a piece of information can be very important in one situation but less important in another situation. What needed to grasp this situational importance of data is a mechanism enabling mining of health data and corresponding situational data.

Main factors for the ubiRank's importance score are Activeness, Relevancy scores. Activeness measures the how active is selected data is in given contextual group. Relevancy score measures how relevant the current record is to situational contexts. Information can be considered as important over certain period of time frame for certain situation. The degree of importance of any record is subject to current condition (situation, time).

Let u denote the user u, r for record r, and c for context c. R denotes the all records space. Then simple ranking of records can be defined as follows:

$$uR(r) = Relevancy(r) + Activeness(r) \qquad (^{8})$$

Activeness. If the record is in active mode if the record is being accessed, modified accessed in certain condition. Activeness of a record is calculated using record usage matrix. If the record is being accessed excessively by doctors and patient himself frequently over some period of time, then it can be said that, record is also active on similar situation in other time frame. The more frequently the record accessed, the more active it gets.

Total activeness score for a record is sum of patient and patient's doctors' usage matrix on that record. Activeness calculation is depicted in Equation **10** and 11.
Let r be a record in record set . Then user *U* be a user in set of r , $T_u$and N_u be user u's total interaction count for record r . Then w_k and $f_k$be weight and frequency of interaction *k* , *k* = {1,2,3,4), each for interaction of viewing, editing, attaching documents, and commenting on health condition record . Then Activeness for given record r is defined as follows.

$$A(r) = \mathrm{EuEr}_{u}{}^{A(\mathrm{r},}\ \mathrm{u}) \qquad (9)$$

$$\mathrm{E\ uET}\ A\,(\mathrm{r}\ \frac{wf,k(u)fk(u)}{N\,f\,(u)} + \frac{-{}^{w}d\ ICOLV\ d(u)}{N\,d(u)}\ {}_{\mathrm{u}} \qquad (10)$$

Weight is calculated as mean frequency of total interactions.

$$Wf^{\mathrm{k}}- \frac{frequencyk}{total\ duration} \quad Wd,k \quad \frac{durationk}{\mathrm{N}totali}$$

Users in this system are not limited to only owner of the data, as data can be shared with doctors and family members. Giving account to possibility of existence multiple users, equation **(11)** can adapt to multiple users. Activeness is strictly dependent on frequency and duration of each usage trial on the record. Activeness score for each user calculated in same manner as owner as in equation 4 and summed to get overall activeness score for record r.

Relevancy Score. Relevancy looks for relevance between records, by analyzing anomalous events in person's life pattern. Relevancy score measures the how this record is relevant to patient's current situation and conditions with respect to his Anomalous activity sequence analysis.

Let R_B be some record created in the past, and denote $R_{AE}(B)as$ set of anomalous activities associated with R_B. Then, again, denote *R„,,„,* be future record to be created and let *RAE (now) =* be sets of anomalous activities associated which *AE (now).* Relevancy score between two records then can be defined as in equation (13).

$$Relevancy\ r\bullet\text{-}\bullet\text{-}\bullet\ Similarity\ (R_{AE}(now), R_{AE}(B)) \tag{12}$$

where, $R_{AE}(B) = tAE_b, AE,,,\}, AE$ E E

R_{AE} $(now) =$ $...,AE_n\},AE$ E $E.$

Then similarity between two records can be found by comparing anomalous activity sets associated with them, Anomalous activity sequence (AAS).

Let $s(i, j)$ be the function to calculate global similarity two AAS, I and J. Then also, let $cop_{11}be$ common ordered pairs in each sequences. Then s is defined as shown in equation (14).

$$SUM = \frac{\qquad\qquad\qquad\qquad}{lnorm} \tag{13}$$

Here, $/_{no}$, is normalized length of sequences, I and J, where each attribute is defined as follows, / = {E_i, E_{i+i}, E_i, , E_n} and j = tE_i, E_{i+i}, E_j, , E_n}.

Let E_x and E_y be activities in activity set of E, and then let ssame(E, , $E_{y)}$ be function to determine if the $E_x$is sufficiently same with E_y. Then ssame(E,E) is defined by equation (15), where m equals total number of situational contexts taken.

$$(14)\ ssame(X,Y) = \{true,\ lE_i \cap E_D I \qquad + 1 \\ false,\ otherwise$$

where, each activityE_i = {L_1, L_2, ..., L_{in}} and e_i E L_i . Equation implies that two activities considered to be sufficiently similar if more than half of the characters are the same, in other words, if total number of same situational contexts those same is greater than the half count, two activities considered to be sufficiently similar.

4 Evaluations on Activity pattern detection

For the purpose of evaluation, this section tests proposed mean life pattern detection and anomaly detection method by applying test data-set. Two types of test conducted with different settings for selection of training data and test data entries. First one is percentage split mode with 66% of data entries used as training data and remainder used as test data set and another is cross-validation with 10-fold. Error rate comparison between two tests is shown in Figure 3. From Figure 3, we can see that test with cross-validation as test data selection method shows slightly more error rates.

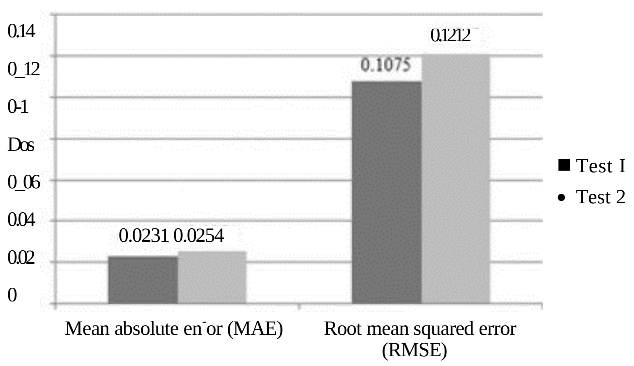


Fig. 4. Error rate comparison

5 Conclusions

In this research, it explored the ranking of health documents and proposed a novel method, ubiRank, short for ubiquitous ranking method, that can be used in everywhere in anytime, as medical emergency could occur at any time. In order to evaluate the proposed ranking method, it needs more testing on test data sets. Development of test environment is in progress currently.

Acknowledgements. This work was supported by the IT R&D program of MKE/KEIT. [10041145, Self-Organized Software-platform (SOS) for welfare devices]

References

1. H. Small. "Co-citation in the scientific literature: A new measure of relationship between two documents". Journal of the American Society for Information Science, 24:256-269, 1973
2. L. Page, S. Brin, R. Motwani, and T. Winograd. "The PageRank citation ranking: Bringing order to the Web." Technical Report, Stanford Digital Library Technologies Project, 1998.
3. G. Jeh and J. Widom. "SimRank: A Measure of structural-Context Similarity",
4. Tversky, A., "Features of Similarity". Psychological Review, 1977. 84(4): p. 327-352.
5. Murata M, Nishimura R, Doi K, Kanamaru T, Torisawa K. "Analysis of the degree of importance of information using newspapers and questionnaires". Proceedings of 2008 IEEE international conference on natural language processing and knowledge engineering (IEEE NLP-KE 2008). 2008. p. 137-144.