Healthone: Personalized Healthcare Recommendation System

Moonsun Shin, Seonmin Hwang, Sungwon Lee, Aeran Jeong, Byungchul Kim

Abstract: In this paper, we propose a healthcare platform, called HealthOne, which is available as an integrated platform for users to manage personalized healthcare applying collective intelligence and ACDT(Ant Colony Decision Tree) based page rank. HealthOne platform provides health-related contents for users in order to manage their own healthcare using smart devices anytime, anywhere according to the PHR profile. To support the personalized recommendation in HealthOne system, ant-colony decision tree and page rank algorithm and machine learning are applied. Furthermore k-means clustering and KNN are adopted for the clustering of similar users based on PHR. We carry out the heuristic experiments of personalized recommendation according to the change of pheromone value.

Index Terms: Personalized Recommendation, Ant Colony Decision Tree Algorithm, Page Rank, Machine Learning,

1 INTRODUCTION
RECENTLY, as the social environment of the healthcare has improved, life expectancy and interest in quality of healthy life have increased. Also, the interest in disease prevention and the needs for health information are increasing. In particular, consumers' needs of health disease information through the Internet are increasing. Therefore, personalized intelligent smart health care is attracting more attention in an aging society. Using PHR(Personal Health Record) people can manage self-health care through day life. In this paper, we design personalized healthcare support system based on machine learning for user-oriented health data management service. The proposed system carries out four main functions such as manipulating PHR/life log data, analysis of user profile, recommendation of the proper health information for each user, and providing inference rules. The advantage of the Health One platform is to search various contents in real time, recommend various healthcare contents such as food or exercise according to the user's inclination, and feed back the results to enable individual self-health care. It can provide users with health-related information crawled periodically by the system according to the user's health condition, and can also provide information processed according to weather and season, and by reflecting the user's feedback information in the recommendation system, can provide good quality information. Personalized recommendation system implements collective intelligent health recommendation service using web crawler, data serialization, and page rank algorithm with ant colony algorithm. In this paper, we implement a collective intelligence-based web / app content service called Health One, and develop a system that provides a personalized recommendation of healthcare information to the appropriate users who are highly interested in healthcare through Health One. A collective intelligence-based personalized algorithm performs the process of web crawling, data serialization, and page ranking with ant colony algorithm. The rest of the paper is organized as follows. Section 2 describes the related research, and Section 3 presents the design of a personalized health care integration support system called HealthOne. Section 4 describes the experiment and evaluation of the implemented algorithm. Finally, Section 5 provides the conclusion and future work.

2 RELATED WORKS
We propose a method of analyzing trends and recommending content based on tag and URL information of users using social web using collective intelligence algorithms and decision trees. Through this, various information reflecting user's preferences is presented to social web users. After collecting data in the form of RDF (Resource Description Framework) consisting of three elements, a predicate, an object, and quantifying the collected RDF tuples using the ant colony optimization algorithm, the pheromones are proposed using the formulas proposed for the user's emotion. The total emotional index for the pheromone values calculated in the previous step is calculated by reflecting the emotional index obtained through SentiWord. In order to verify the validity of the proposed method, we show that the user's sentiment trend, which is calculated based on the overall sentiment index, is properly analyzed through comparison with the real life of the user. PageRank algorithm is implemented and used in commercial search engines, but for the commercial reasons, the results of the implementation techniques are hardly published. This paper describes the implementation techniques of the PageRank algorithm and suggests the input / output data structures and four major implementation techniques for applying them. This paper shows how to apply the PageRank algorithm as an example of system that calculates the PageRank value of a real web document.

2.1 Serialization of Data
We preprocessed self-diagnostic data to serialized data for the efficient processing. Serialization means that the user's checkup response items are preprocessed into numbers and composed into a single string. Each index value of serialized data is referred to extracting a category of content to be provided for the user. For example, if the diagnostic data for gender and age are 'male' and '40s', then the value of the index will be 0 and 2 in the serialized data, respectively. Using this, it is determined whether the health content to be provided to the user is appropriate or not. The self-diagnostic data provided by users can be preprocessed according to the referring table shown in table 1.

Serialization example of self-diagnostic data:
001010203102410311202001000000

---

*Moonsun Shin, ¹Dept. of Computer Engineering, Konkuk University, Chungju 27478, Korea, E-mail: msshin@kkku.ac.kr*

*Seonmin Hwang, E-mail: smhwang@kkku.ac.kr, Corresponding author*

*Sungwon Lee, ²Dept. of Information Statistics, Chungbuk Nat’l University, Cheongju, Korea, E-mail: rhsw3@hanmail.com*

*Aeran Jeong, E-mail: jskim@kkku.ac.kr*

*Byungchul Kim, ³Dept. of ICT, Baekseok University, Cheonan, Korea, E-mail: bckim@bu.ac.kr*
2.2 Collective Intelligence based Personalized Recommendation

(1) Extend ant colony algorithm
Page rank refers to a method of weighting a document having a hyperlink structure such as the world wide web according to relative importance. It is the most representative algorithm for impact analysis among the centrality indices that are widely used among network analysis techniques. Collective intelligence refers to the outcome or process itself obtained by gathering a large number of people into a cluster and performing the collective activities and roles, and is used to derive insights by combining the members' behaviors, choices, and ideas. The ant colony algorithm is one of these algorithms that uses the collective intelligence to model the process of ants’ goal. The pheromone evaporation and accumulation occur according to the movement of each ant, and the pheromone is the best choice when there are several paths. Based on the accumulated pheromone value, the content is judged to be valuable by various users.

The following equation is from the ant colony algorithm.

\[ \tau_{ij} = (1 - \rho) \times \tau_{ij} + \sum_{k=1}^{n} \Delta \tau_{ij} \]

\[ \Delta \tau_{ij} = \frac{Q}{\text{ant's movement of point } i \text{ to } j} \]

Extending ant colony algorithm to apply to personalized recommendation in HealthOne platform, the ant and the pheromone value left by the ant were set as the number of views and satisfaction left by the user. The contents that show high views and satisfaction in the category maintain high pheromone values, while other content that does not have a pheromone value continuously evaporates and converges to zero. Since the pheromone value determines the recommendation level of the content, the radical change decreases the reliability. To avoid such radical changes, evaporation rates must be determined experimentally in a heuristic manner. It is needed to rank each content by category to provide appropriate content to users. The user is provided with higher quality content by ranking the pheromone value, which is updated with the satisfaction indication for determining the value of the provided contents. It also provides random content because it may be of exceptional value and provides up-to-date health information content that may have low pheromone values. The latest two contents can be selected for the page ranking and also five contents with high pheromone value based on the date information while crawling. In addition, two types of contents are randomly selected and provided to the user to maintain freshness of pheromone value.

(2) Web crawler
Web crawler provides the ability to automatically collect web documents of a specific site from the web which consists of a large amount of web documents. It traverses web servers and gathers information from various webpages automatically and analyzes the contents of web pages. The information collected by web crawler includes the URL, title, and metadata of the page, and additionally counts the number of keywords included in the page. We propose a distributed web crawler that solves the problems of existing web crawlers through the RSS and Google Search APIs of the website, and utilizes RMI and NIO to minimize the network connection between server and client to provide fast crawling. In addition, the proposed web crawler provides a function of automatically extracting only important contents used for analysis by comparing keyword similarities with respect to tags constituting a web document. Finally, the performance evaluation results of the existing web crawler and the proposed crawler demonstrate the superiority of the proposed web crawler.

3 HEALTHONE SYSTEM ARCHITECTURE

3.1 Workflow of HealthOne system
The planning of the system stores the user's health information data according to the survey items prepared in advance, and the system crawls and stores the health information for each predetermined category at regular intervals in advance. When the user accesses, it is provided with three types of health information, which are newest, high preference, and random, according to each user's health information.

<table>
<thead>
<tr>
<th>index</th>
<th>name</th>
<th>Desc</th>
<th>Valid value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0~9</td>
<td>IDX no.</td>
<td>id</td>
<td>00000000000-9999999999</td>
</tr>
<tr>
<td>10</td>
<td>sex</td>
<td>0: M, 1: F</td>
<td>0.1</td>
</tr>
<tr>
<td>11</td>
<td>age</td>
<td>0: age&lt;=19 1: 20&lt;=age&lt;40 2:40&lt;=age&lt;70 3: age&gt;=70</td>
<td>0~3</td>
</tr>
<tr>
<td>12</td>
<td>meals</td>
<td>Meals of a day</td>
<td>1~4</td>
</tr>
<tr>
<td>13</td>
<td>Regular meals</td>
<td>0: N, 1: Y</td>
<td>0.1</td>
</tr>
<tr>
<td>14</td>
<td>Night meals</td>
<td>0: N, 1: Y</td>
<td>0.1</td>
</tr>
<tr>
<td>15</td>
<td>Count of alcohol</td>
<td>0: N, 1-7: Y</td>
<td>0~7</td>
</tr>
<tr>
<td>16</td>
<td>Amount of alcohol</td>
<td>0: N, 1: Y</td>
<td>0~1</td>
</tr>
<tr>
<td>17</td>
<td>smoking</td>
<td>0: N, 1: stop, 2: Y</td>
<td>0~2</td>
</tr>
<tr>
<td>18~19</td>
<td>Cardiovascular exercise</td>
<td>Total hour of exercise for a week</td>
<td>00~99</td>
</tr>
<tr>
<td>20~21</td>
<td>Anaerobic exercise</td>
<td>Total hour of exercise for a week</td>
<td>00~99</td>
</tr>
<tr>
<td>22</td>
<td>High blood pressure</td>
<td>0: N, 1: Y</td>
<td>0.1</td>
</tr>
<tr>
<td>23</td>
<td>diabetes</td>
<td>0: N, 1: Y</td>
<td>0.1</td>
</tr>
<tr>
<td>24</td>
<td>Liver cancer</td>
<td>0: N, 1: Y</td>
<td>0.1</td>
</tr>
<tr>
<td>25</td>
<td>Stomach cancer</td>
<td>0: N, 1: Y</td>
<td>0.1</td>
</tr>
<tr>
<td>26</td>
<td>Lung cancer</td>
<td>0: N, 1: Y</td>
<td>0.1</td>
</tr>
<tr>
<td>27</td>
<td>Thyroid cancer</td>
<td>0: N, 1: Y</td>
<td>0.1</td>
</tr>
<tr>
<td>28</td>
<td>Breast cancer</td>
<td>0: N, 1: Y</td>
<td>0.1</td>
</tr>
<tr>
<td>29</td>
<td>Other cancer</td>
<td>0: N, 1: Y</td>
<td>0.1</td>
</tr>
</tbody>
</table>
For the information appropriate to the user's health, the user selects a category of content to be provided to the user by referring to an index value generated by serializing the user's self-diagnosis item and provides crawling information corresponding to the category. In this case, information of high suitability is preferentially provided by the page rank algorithm to which the ant colony algorithm is applied. Figure 1 shows the overall process of the HealthOne system. The web crawler periodically crawls using the given keywords. Then these data about crawled content have been stored in MongoDB. The keywords given are the keywords defined in advance in conjunction with the self-health diagnosis item. Administrator can add or delete keywords arbitrarily. When crawling, the black list items could be referred, which were added to the dictionary to prevent unnecessary content storage. The categorization of the keyword set to the crawled content must be added and stored in MongoDB. For example, if the content is crawled with keywords such as "overeating" or "irregular meals," it is grouped into large categories such as "eating habits." This is a standard filtering of content when the user does not want the corresponding category.

The algorithm described above is executed based on the accumulated contents. Quality content, selected as a result of collective intelligence, is ranked in each category and maintained in the database. This ranking is flexible according to user preferences. The user enters self-health diagnosis information through a web or an app. At this time, data related to the user has been stored in the MySQL database, and serialized self-test information can be stored in the MongoDB. These two databases are used to take advantage of the stability of data storage and the speed benefits of the unstructured database. When a user enters a self-health check input, the system analyzes each index of the serialized data and provides a web/app with a page ranked in the corresponding category. When the user indicates satisfaction or preference for the content, the recommendation module calculates the pheromone value for the content using the equation mentioned above. After that, the page rank module operates to update the knowledge base once more.

3.2 Software Architecture of HealthOne

The HealthOne system consists of four components such as a page ranking module, a recommendation module, a gathering information module and an administrator module. The software architecture of the HealthOne system is shown in figure 2. The page rank module maintains valuable information continuously and storing the collected contents in knowledge base database, user self-diagnosis data and current weather data. It consists of a recommendation module that receives and displays health information contents from the knowledge base and an administrator module that manages the entire system. The gathering information module performs a function of crawling and storing health information content on the web or collecting weather information through a weather API. The page rank module processes the collected contents and stores them in the knowledge base to maintain valuable information continuously.

4 IMPLEMENTATION AND EXPERIMENTS

4.1 Implementation

The software we used to implement HealthOne system were shown in table 2. HealthOne server was constructed by apache, node.js and PHP. Web crawler was implemented using Python. Android app was developed in Android studio and java. Both formal database and nosql database were used.
For efficient management of database, we constructed two databases. The mongoDB used as a typical nosql database, which had a schema-less structure, that meant any type of data could be stored and high performance could be provided in the operations of read/write. It also could be stored as json format, which might be intuitive and easy to develop. However, since join operation does not exist, it should be designed so that join is not necessary when structuring data, and there existed a problem that performance had been decreased as the size of B tree has been grown.

<table>
<thead>
<tr>
<th>OS</th>
<th>Linux Ubuntu 16.04 LTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Programs</td>
<td>Node.js / Python3 / Apache2 / PHP7 / Java / Android studio</td>
</tr>
<tr>
<td>Database</td>
<td>MongoDB / Mysql</td>
</tr>
</tbody>
</table>

Considering the efficiency of the whole database, both two types databases were used. The user information was stored in mysql and the health information was stored in mongoDB. HealthOne Web application is shown in figure 3 and android app is shown in figure 4. The web application was developed responsive so users can easily access in mobile device without installing app client. HealthOne app provides pedometer, heart rate check, and food information check for systematic self-healthcare. Pedometers and heart rate checks provide instant notification of the current status of user based on measured results. Pedometer has a mechanism that works by detecting events using the gyro sensor. Heart rate measurement works by detecting events through the heart rate sensor on the back of Android via the Google Fit API, and the criteria for heart rate ranges are based on commonly used numbers.

4.2 Experiments
The update formula of a pheromone, which is modified for the HealthOne system by applying the likelihood and the hits is shown as follows.

\[
\tau_k = (1 - \rho) \times \tau_0 + \left( \frac{|LL|}{|CEL|} + \frac{|LH|}{|CEH|} \right)
\]

\[
\tau_{k+1} = (1 - \rho) \times \tau_k + \left( \frac{|LL|}{|CEL|} + \frac{|LH|}{|CEH|} \right)
\]

Where \( T_0 \) is the initial value, \( \rho \) is the evaporation rate, \(| LL |\) is the Link Like of the link, \(| CEL |\) is the Category Entire Like, and \(| LH |\) is the Link Hits. The denominator \(| CLH |\) means a Category Link Hit of links of the same category as the link. Figure 5 showed the pheromone change according to the preference display using random data. The horizontal axis represents the preference and the vertical axis represents the pheromone value.

Since the pheromone value determines the recommendation level of the content, the radical change decreases the reliability. To avoid this radical change, the evaporation rate must be found to be insensitive. Through the heuristic experiment, we experimented to find the appropriate constant and apply it to the equation. The algorithm was applied to HealthOne using the obtained values. The initial value of pheromone \( T_0 \) was 1.0, the evaporation rate \( \rho \) was 0.05 were obtained as suitable results in the experiments. Correlation analysis was performed for each variable of the users' self-diagnosis data. As a result, the relationship between physical condition, lifestyle, and disease was figured out. To cluster similar users, the K-means clustering was carried out, and as a result the three cluster were generated labeled normal,
dangerous, and high risk respectively. Figure 6 showed the experimental results of K-means clustering.

![Image](image.png)

**Fig. 6.** Change of pheromone according to the preference

Through the experiments, it was verified HealthOne performed to classify similar users according to their PHR and provide personalized recommendation of health information contents applying collective intelligence and ACDT based page rank.

5 CONCLUSIONS

In this paper, we proposed a healthcare platform called HealthOne, which could be used as an integrated platform for users to manage personalized healthcare. It also performed to support a PHR-based intelligent healthcare contents applying collective intelligence and ACDT based page rank. We designed software architecture of the HealthOne platform and implemented those algorithms we extended. We extended the update formula of a pheromone, which was modified for the HealthOne system by applying the likelihood and the hits in order to provide users appropriate healthcare contents. We have verified the usefulness of the proposed system through the experiments of the algorithms we applied and extended. We also have developed the app client of the HealthOne. As a future work, we are going to research machine learning in order to advance the HealthOne system and investigate the data stored in HealthOne and provide useful knowledge by applying big data analysis.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A3B03033944).

REFERENCES