IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 12, 2021, accepted April 6, 2021, date of publication April 16, 2021, date of current version April 29, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3073775

Multi-Level Health Knowledge Mining Process in P2P Edge Network

JI-WON BAEK[1] **AND KYUNGYONG CHUNG**[2]
[1]Department of Computer Science, Kyonggi University, Suwon-si 16227, South Korea
[2]Division of AI Computer Science and Engineering, Kyonggi University, Suwon-si 16227, South Korea

Corresponding author: Kyungyong Chung (dragonhci@gmail.com)

This work is supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant 21CTAP-C157011-02).

ABSTRACT Chronic diseases are increasing due to westernized eating habits and everyday life changes, and healthcare and disease prevention should be managed based on constant interest. Users, who are not health professionals, have difficulty in obtaining accurate information related to healthcare due to noise problems such as subjective opinions, distorted information, and exaggerated information. There is a need for a method that enables users to obtain meaningful information for healthcare and disease prevention in real-time among the vast amounts of data collected through search. In this study, we propose a multi-level health knowledge mining process in a P2P edge network. The proposed method suggests a P2P edge network to solve the overload problem of P2P networking, the noise problem, and the security problem of cloud computing and mines the health knowledge through the mutual information according to the association rules. In addition, the results of health knowledge mining are visualized to propose a method by which users can easily receive relevant health information. As a result of the performance evaluation, the F-measure using recall and precision is 83%, 79%, 75%, 74%, and 73% of the support ratings of 10%, 20%, 30%, 40%, and 50% Was derived. Accordingly, it is possible to process and analyze healthcare-related information in real-time through a multi-level based health knowledge tree based on the association of data collected by P2P edge computing. In addition, by visualizing meaningful information to the user through the embedding network structure, it provides personalized information for intuitive understanding.

INDEX TERMS P2P edge network, data mining, multi-level, emerging health knowledge, hybrid P2P.

I. INTRODUCTION

In modern society, the burden of medical expenses is increasing due to an increase in single-person households, an increase in the number of elderly people, and an increase in the number of people with chronic diseases. Healthcare combined with the Internet and medical technology is developing, and interest in healthcare is increasing. The number of people with chronic diseases is constantly increasing due to the surrounding environment, genetic factors, and westernized living habits [1]–[3]. Chronic diseases are diseases that require constant management such as hypertension, diabetes, arthritis, and rhinitis, and life care is required to prevent symptoms from worsening. In addition, efficient healthcare services with low medical expenses should be provided [4]–[6]. Health and disease may be temporary in lifestyle habits and are divided into stages of causes, consequences, and complications that may worsen. For example, the national health information portal of the Korea Centers for Disease Control and Prevention has suggested that the result of hypertension may be caused by cause factors such as age, family history, obesity, stress, and lack of activity. Also, complications of hypertension include hemorrhagic stroke, heart failure, and myocardial infarction [7], [8]. Therefore, continuous healthcare and disease prevention require attention and efforts to prevent the complication risk of chronic diseases in daily life.

The development of edge computing enables us to share information available anywhere, and in hybrid P2P networking that connects distributed wired and wireless networks [9]. Recently, high-performance processes such as virtual reality, big data processing, machine learning, and artificial intelligence are distributed and processed, enabling efficient data processing. It is also possible to update, share,

The associate editor coordinating the review of this manuscript and approving it for publication was Sudhakar Babu Thanikanti.

and extend the data provided by various IoT devices and data in social networks, requiring the technology to process data in real-time [10]. Knowledge is used in various fields such as marketing, medical and health industries through distributed processing and meaningful information analysis in heterogeneous big data. In order to collect data from news, blogs, SNS, web applications, and process the data, there is a need for dispersion and parallel technologies that distribute the big data across multiple servers and collect them on each server to summarize the result. The Map-Reduce process, presented by Google as open-source software, processes big data at high speed with a Map step to process big data distribution and a Reduce step to collect intermediate results and eliminate redundancy. Therefore, there is a limit to processing on one server, so the edge computing technology that introduces the concept of distributed processing is highlighted as the element technology of the big data platform together with Hadoop distributed processing of actually commercialized Oracle, Microsoft, Samsung, Dell and the like.

In a P2P network, it is difficult to obtain desired meaningful information from a large amount of data due to noise such as subjective opinions, distorted information, and exaggerated information in the data collected by the user. For example, when searching for information related to healthcare, general users without medical knowledge have difficulty in making accurate and reliable information decisions. Therefore, it is necessary to provide the user with health information such as reliable and meaningful health and disease, prevention according to symptoms, management, exercise method, and diet in an easy-to-understand manner. With the development of smartphone applications and wearable devices using IT convergence technology in the healthcare field, technology for providing personalized health management in real-time to users has been advanced [3], [4], [11], [12]. The smartphone application developed by Samsung Health [13] provides life-care services by recording and analyzing exercise and activity logs in P2P networking in consideration of user's convenience. In addition, step counters, sleep pattern recording, and diet management functions provide useful information for physical strength enhancement and diet. Also, wearable devices developed by the InBody Band [14] check the muscle mass, body fat percentage, and momentum in real-time to provide information about the user's health status in real-time. It calculates the number of steps for 24 hours and calories and automatically recognizes and analyzes sleeping hours and health status. It also provides healthcare services to users by calculating daily recommended calorie intake and momentum according to individual BMI values [14]. As shown above, the element technology is needed to maintain the security of medical information and provide reliable information in real-time in the P2P edge network. In this study, we propose a multi-level health knowledge mining process in a P2P edge network. The proposed method proceeds as follows:

- First, requests data from a shareable P2P network between nodes connected independently without a central server, such as documents and memory. As the number of nodes increases, an overload problem and noise problem occurs in the P2P network. It is also vulnerable to security because it uses the Internet. To solve this problem, medical information is collected and processed in the P2P edge network. The edge network can strengthen medical security through distributed processing, and solve the overload problem by distributing the operation of data. The knowledge generated in the edge network is stored in the cloud computing repository for reuse and expansion.
- Second, the collected health information generates Bag of Health Word through preprocessing and morpheme analysis. In a health corpus, the frequency and distribution of words are used to reconstruct them as health transactions. Apriori algorithm of data mining and multi-level association rules are used to construct a health knowledge tree. The multi-level association rules extend the meaning and find the association rules of the words constituting the candidate item set as well as the frequent items.
- Lastly, the association between words through mutual information (MI) is used to generate health knowledge. Based on the generated multi-level knowledge tree and MI, the results expressed as the embedding network structure are visualized to provide the user with health information and disease prevention information.

This study is organized as follows: Chapter 2 describes the related researches of health-based P2P edge computing, Chapter 3 describes the proposed multi-level health knowledge mining process in the P2P edge network, Chapter 4 describes the performance evaluation, and Chapter 5 provides a conclusion.

II. HEALTH BASED P2P EDGE COMPUTING

In modern society, demand for smart devices such as IoT and wearable is increasing, and the use of the Internet is being activated anytime and anywhere. Using IoT devices, users can collect, share and communicate vast amounts of data scattered from SNS in real time. For this reason, edge computing is used to reduce the time to collect and process vast amounts of data [15]. In addition, as healthcare technology becomes more sophisticated, users receive appropriate healthcare in hybrid P2P networks regardless of time and places [16], [17]. In a P2P network, resources can be shared between independently connected nodes without a central server. However, if the number of nodes increases, an overload problem occurs. When data is transferred from P2P, the cloud server processes storage, processing, and content usage. Cloud computing can reduce money, manpower and data loss to build a server in a hybrid P2P network [18]. Also, if the Internet is available, data stored on the server can be used regardless of time and places. On the other hand,

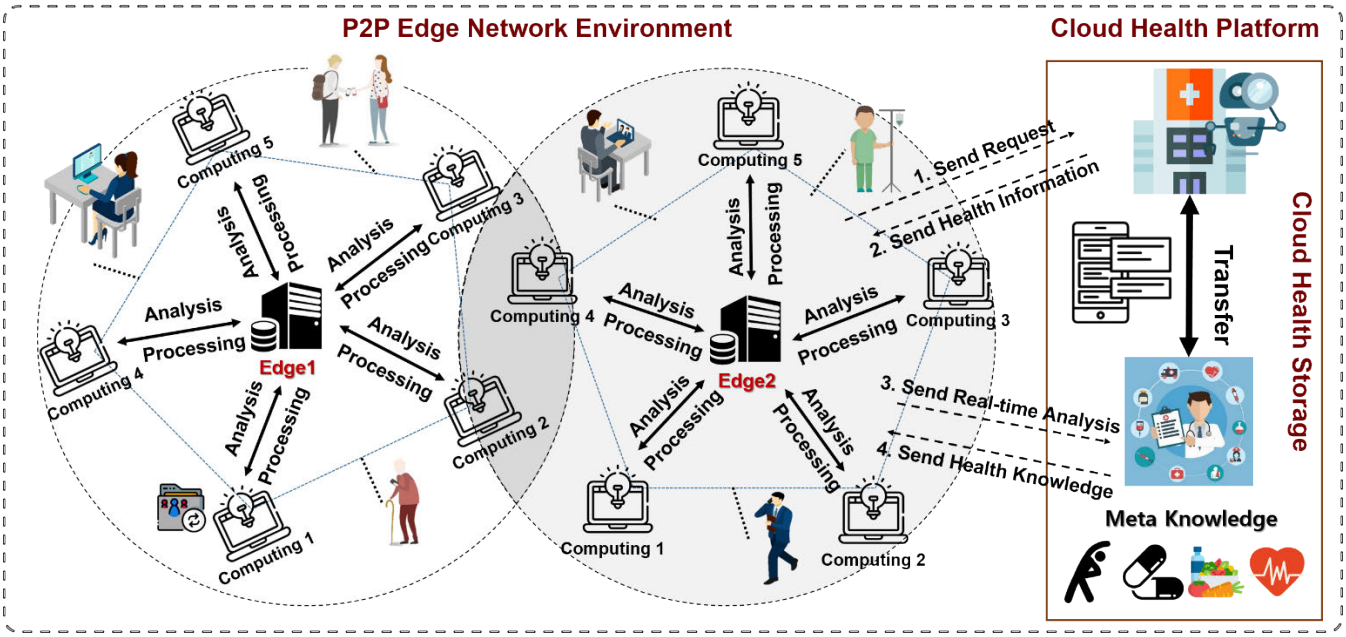

FIGURE 1. Block diagram of health-based P2P edge computing.

the cloud has a security problem of data leakage if the server is hacked. Edge computing can process vast amounts of data in real time through small servers that are distributed at the edge of the network [15], [19]. Unlike cloud computing, this is located in a place where users' IoT devices and computing for the computation of big data are relatively close. Also, since it is a distributed system, it can escape from the concentration target of hacking attack. As a result, it has the advantages that it can process data generated in real-time in a short distance, can reduce the bandwidth required for Internet use and big data processing time and can enhance security [19]. Hybrid P2P can share and exchange data between individuals and individuals, individuals and medical institutions connected independently via the Internet [18]. It processes and provides the health status in real-time on edge computing's health platform through IoT devices. Figure 1 shows the block diagram of health-based P2P edge computing.

In Figure 1, the edge processes the big data before it is sent to the cloud repository and provides the user with the knowledge through the analysis. The user and server are connected by hybrid P2P networking. If the user transmits information to the server, the server analyzes and processes it in real time, and then provides the feedback of the information to the user again. The knowledge analyzed in the health server is stored as meta knowledge in the cloud repository for reuse and expansion. Meta knowledge is used to express knowledge, and as a result, knowledge provides a method of decision-making that can solve problems [18]. For example, meta knowledge that expresses knowledge about healthcare and disease prevention includes exercise, nutritional diet, and so on. Chen *et al.* [19] proposed a healthcare system based on edge cognitive computing. The proposed method can monitor and analyze the user's physical health using cognitive computing and adjust the resource allocation of the edge-computing network according to the health risk level. This is a system that improves the patient's survival rate in an emergency, optimizes computing resources and improves the user's environment. Li *et al.* [20] proposed a block-chain technique for P2P cloud repository. The proposed method divides the file into encrypted data chunks and applies a block-chain-based security technology for distributed cloud storage that can be arbitrarily uploaded to a node in a P2P network. This is a technology that provides more secure and reliable cloud storage to companies or individual users.

III. MULTI-LEVEL HEALTH KNOWLEDGE MINING PROCESS IN P2P EDGE NETWORK

A. COMPOSING BAG-OF-HEALTH WORDS IN P2P EDGE NETWORK

Bag-of-health words collect health data related to healthcare and disease prevention in the P2P edge network. It uses hybrid P2P networking in a health platform that connects users and healthcare servers [17], [18]. Health data uses IoT devices to request the data to analyze, share, manage, and prevent health conditions to the health platform. The edge network provides the health information by processing it in a distributed server located at the edge [15], [19]. It also generates and provides knowledge as meta knowledge by analyzing the data transmitted and requested by the user in real-time. Meta knowledge provided is knowledge of complications, symptoms, causes, results, and management of diseases. This is text-format data, which extracts words in the form of nouns through preprocessing. Preprocessing eliminates unnecessary elements through the processes of elimination of stop words,

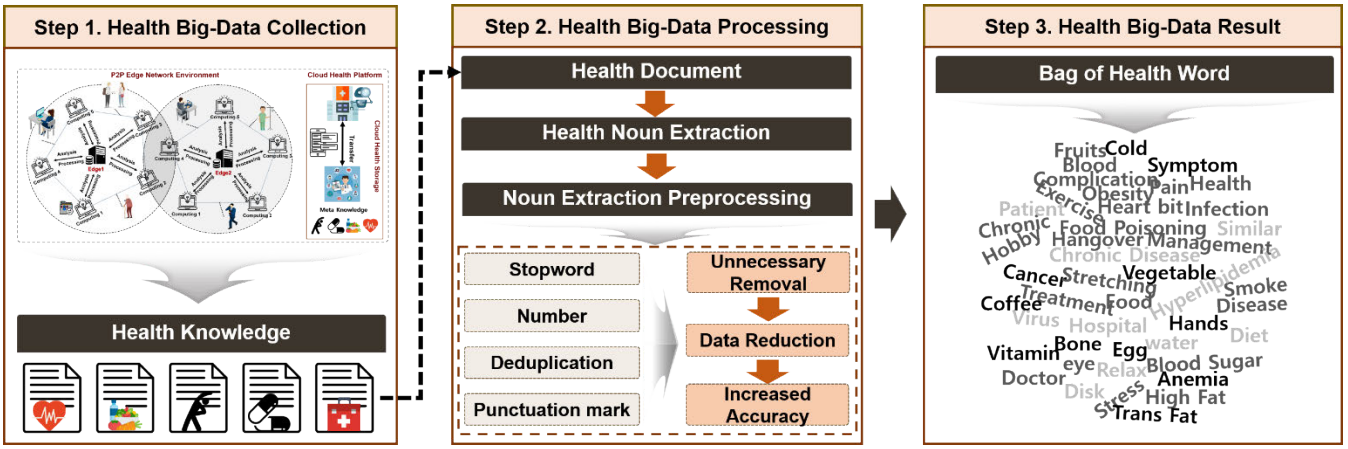

FIGURE 2. Configuring process of bag-of-health-words in P2P edge network.

elimination of punctuation marks, elimination of numbers, and elimination of duplication. It also reduces the size of the data and improves the accuracy of the analysis. Elimination of stop words dispose of meaningless words such as postpositions, pronouns, and conjunctions. The elimination of punctuation marks handles delimiters to distinguish sentences. The elimination of number handles dummy information indicating a number. The elimination of duplication handles words and sentences that appear in duplicate in one sentence. After preprocessing, Bag-of-health-words composed of words extracted from health information is constructed. Figure 2 shows the configuring process of bag-of-health-words in a P2P edge network.

B. DISCOVERY OF ASSOCIATION RELATION USING KNOWLEDGE MINING

The collected health information generates bag-of-health-words through preprocessing and morpheme analysis. In health corpus, the frequency and distribution of words are used to reconstruct them as a health transaction. The health transaction is used to find the association using knowledge mining. The Apriori algorithm and multi-level association rules are used to find the association in the health transaction. The Apriori algorithm is a method to extract frequent items from the health transaction and find rules in which the association among independent other items meets the minimum support [5], [6]. The minimum support set through repetitive performance evaluation is used for knowledge mining to find meaningful knowledge [21], [22]. Figure 3 shows the discovery process of association rules using Apriori algorithm.

In Figure 3, *Apriori_Gen()* function of Apriori algorithm is used to explore a frequent item set of the transaction with a minimum support of 3. The *Apriori_Gen()* function generates a candidate item set to explore a frequent item set and consists of a Join step and a Prune step to reduce the number of candidate item sets. It uses the health transaction as input. The Join step of creating n + 1 candidate item sets is carried out in a frequent item set composed of n candidate items.

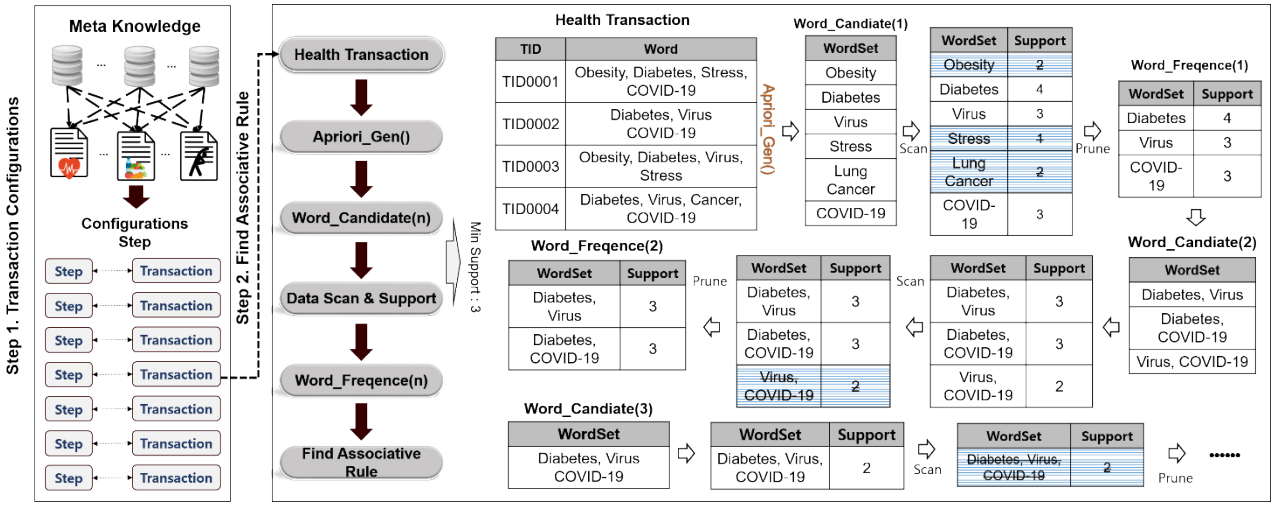

FIGURE 3. Discovery process of association rules using Apriori algorithm.

If a subset of n + 1 candidate item sets does not exist in the transaction and does not satisfy the minimum support, the subset is deleted to conduct the Prune step to create a candidate item. For example, the item with a minimum support of 3 is scanned in the candidate item set of {Obesity, Diabetes, Virus, Stress, Cancer, and COVID} with $n = 1$. The Prune step to remove items {Obesity, Stress, and Cancer} that do not satisfy the minimum support is carried out construct frequent item sets {Diabetes}, {Virus}, {COVID-19} with $n = 1$. The Join step to construct candidate item sets with $n = 2$ in {Diabetes}, {Virus}, {COVID-19} is conducted to create {Diabetes, Virus}, {Diabetes, COVID-19}, {Virus, COVID-19}.

As shown above, the creation of candidate item sets, minimum support scan, prune step, the generation of frequent item sets, and join step are repeated. *Apriori_Gen()* stops searching for frequent item sets if there are no item sets that can be composed in a transaction. *Word_Candidate(n)* means a candidate set consisting of n words. The generated candidate item set is scanned to generate a frequent item set by sorting the set that satisfies the minimum support. *Word_Frequency(n)* means the frequent item set generated. The support and reliability are specified from the frequent item set to find the association rules between words [23]. The support is the rate for finding words that appear frequently in a transaction. The reliability means the assumption's accuracy for the conclusion, and the higher the reliability, the higher the association. Lift is a criterion for evaluation of association rules and shows the correlation between condition and result. Lift means a negative correlation when it is less than 1, and a positive correlation when it is greater than 1. In this study, we extract a frequent item set whose candidate item set satisfies a minimum support of 10% or more in a transaction [24]. In the health transaction, the Apriori algorithm is used to create association rules. Table 1 shows the extracted association rules from frequent item-sets.

TABLE 1. Extracted association rules from frequent item-sets.

No	Health Association Rules	Support	Confidence	Lift
1	{Obesity}→{Hypertension}	0.124	1.000	2.059
2	{High fat, Chronic Disease}→ {Trans fat, Cancer}	0.118	1.000	2.333
3	{Anemia, Diet, Cure}→ {Light head, Vitamin, Fruit}	0.100	0.875	1.494
4	{Disk, Pain, Cause}→{Position}	0.151	1.000	2.059
5	{Food, Virus}→ {Food Poisoning}	0.127	1.000	2.059
6	{Obesity, Complications}→ {Diabetes, Hyperlipidemia}	0.132	1.000	2.333
7	{Disk, Stretching}→ {Neck, Waist, Shoulder}	0.302	1.000	1.707
8	{Virus}→{Infection}	0.129	0.900	2.172
9	{Stress, Management}→ {Exercise, Sleep, Hobby, Relax}	0.129	0.900	1.537
10	{Hangover, Cure}→ {Water, Egg, Exercise, Coffee}	0.114	0.800	2.074
11	{Cancer, Ablation}→ {Operation, Radiation, Drug}	0.154	0.800	1.647
12	{Smoking, Cancer}→ { Laryngeal, Lug, Bladder}	0.129	0.900	1.537
13	{Fast-food, Overeating}→ {Cause, Complication, Obesity}	0.171	0.857	1.463
…	…	…	…	…

In Table 1, the association rules of {Obesity, Complications}→{Diabetes, Hyperlipidemia} are support 13.2%, reliability 100%, lift 2.333, and the evaluation scale is composed. This means that there is a very high and positive correlation between words composed of {Obesity, Complications, Diabetes, Hyperlipidemia}. In addition, the health and disease information provided by the Korea Centers for Disease Control and Prevention shows that the complications of obesity are related to diabetes and hyperlipidemia. In fact, there are diagnosis results that the incidence of diabetes and cardiovascular diseases is increased if there are metabolic syndromes such as a rise in blood glucose, blood lipid abnormality, an increase in body fat and elevation of blood pressure.

C. HEALTH KNOWLEDGE TREE USING MULTI-LEVEL ASSOCIATION RELATION

The health knowledge tree consists of multi-level association rules. This is generated based on the MI and association of candidate items constituting frequent items of health words. The Multi-Level Association Rule finds association rules by applying concept hierarchy classified by reflecting knowledge in advance. The concept hierarchy discovers the conceptual features of a word and expresses the meaning of the word hierarchically. This extends the meaning to calculate and express the similarity between words. The Multi-Level Association Rule uses the support and reliability, which are the evaluation scales of the Apriori algorithm and finds the relationship between candidate items. As a depth search method based on Apriori algorithm, it finds a rule that satisfies the forms of upper level-middle level-lower level [25], [26]. In addition, it is possible to analyze the hierarchical level of a pattern hidden in different items by evaluating the hierarchical structure for data items [27]–[29]. For example, hidden association rules are found such as {Obesity, Complication, Fast Food, Chicken, Overeating, Obesity, Complications, Diabetes, Hyperlipidemia} by expanding item sets in order to find the relationship between {Obesity, Complications}. Figure 4 shows the multi-level association rules based health knowledge tree. The health knowledge tree in Figure 4 shows negative lifestyle for the reason of obesity-related complications. It also shows that positive habits such as exercise among life habits can prevent obesity. The hierarchical tree can be used to identify the association between words by searching for various paths. In order to find out the degree of the association, we find the mutual association using the MI

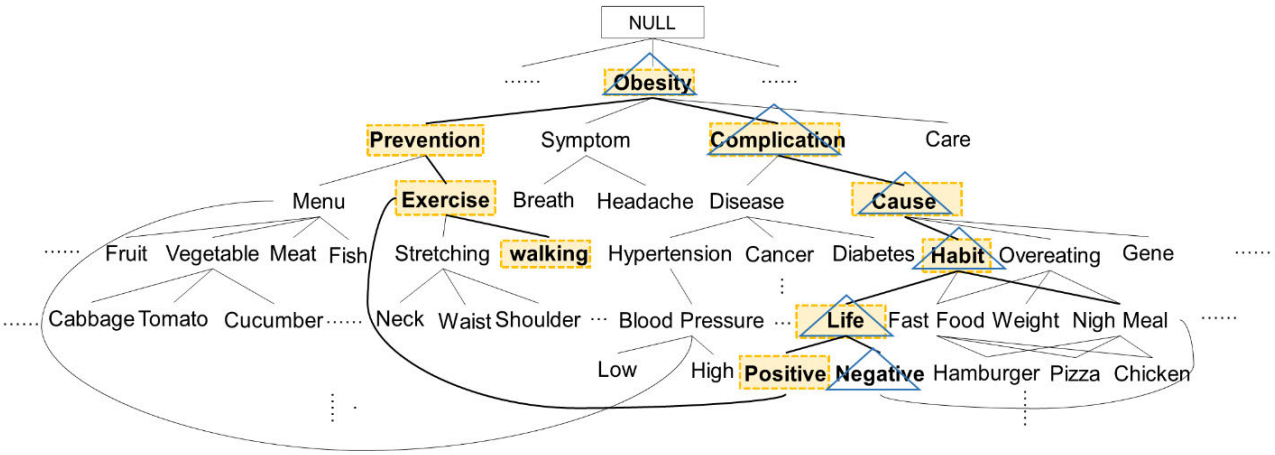

FIGURE 4. Multi-level association rules based health knowledge tree.

based on the health knowledge tree. The *mutual information (MI)* is a method of finding the information amount of the mutual association about how closely the two words are related [30].

This is based on the probability of each frequency of words and the frequency probability of words appearing at the same time. Co-Occurrence means the number of appearance of different words at the same time. The larger the number, the more closely related the two words are, and it is used to measure the similarity between words [31]. In the information theory, the amount of information uses the probability value that the event occurred and the value taking a log in a reciprocal [25], [26], [30]. Equation (1) represents the mutual information.

$$MI\ (w_1, w_2) = \log_2 \frac{f(w_1, w_2)}{f(w_1) \times f(w_2)} \tag{1}$$

w_1 and w_2 mean the targets to be compared. $MI(w_1,w_2)$ represents the mutual information between w_1 and w_2. $f(w_1)$, $f(w_2)$, and $f(w_1,w_2)$ refer to the probability of frequency of w_1, the probability of frequency of w_2, and frequency of simultaneous occurrence of w_1 and w_2, respectively. If the result value of the *MI* is greater than 0, the two words are relevant. If it is less than 0, it means that there is no relevance. In this study, we apply the lift that can evaluate the association rule by Apriori algorithm to the *MI*. This is to judge the degree of association between words by using the MI between words. w_1 and w_2 are an independent relationship, and the lift represents the frequency at which word w_1 and word w_2 appear at the same time. Taking a log in the lift value equals the mutual information value. The lift based mutual information according to the association mining is defined as Equation (2).

$$\begin{aligned} \text{Lift}\,(w_1 \rightarrow w_2) &= \frac{P\,(w_1 \cap w_2)}{P\,(w_1) \times P\,(w_2)} \\ MI\ (w_1, w_2) &= \log_2 \text{Lift}\,(w_1 \rightarrow w_2) \end{aligned} \tag{2}$$

Time complexity means the time taken for an algorithm to solve a problem or the count of operations in the algorithm. Table 2 shows the result of the comparison between Apriori and FP-Tree based association rule algorithms and the proposed improvement based mutual information method in terms of time complexity. The comparison is made after the data scan count, whether to create candidate items, and time complexity is designed.

TABLE 2. The result of the comparison between Apriori and FP-Tree based association rule algorithms and the proposed improvement based mutual information method in terms of time complexity.

	Apriori	FP-Tree	Ours
Number of Data Scan	C_n+1	2	C_n+1
Create Candidate Item	Yes	No	Yes
Time Complexity	$O(2^n)$	$O(n\times(2^m\text{-}1))$	$\log_2 O(2^n)$

As shown in table 2, the Apriori algorithm creates a candidate set of $a(a+1)/2$ in the first search step in the condition where the number of frequent items is an infrequent pattern search. In the next step, it also creates candidate items, and the execution time of association rules is different depending on the minimum support count and transaction count. Accordingly, in the worst case, the time complexity is $O(2^n)$. The proposed improvement based mutual information method applies log_2 to an improvement on the basis of the Apriori association rule. Therefore, log_2 is applied to the time complexity of the Apriori association rule, and a calculation is made with $log_2 O(2^n)$. In FP-Tree, m represents the number of nodes in which the total of all paths for finding frequent items for a certain item meets the minimum support. In addition, Apriori and the proposed improvement based mutual information method generate a candidate set, whereas FP-Tree doesn't do so. Therefore, in FP-Tree, time complexity continues to increase along with the value of m. Table 3 shows the results of the lift-based mutual information according to

the association rules. In this study, if the value of Lift, which is an evaluation measure of association rules, is greater than 1, it is judged to be a meaningful relation and used.

TABLE 3. Results of the lift-based mutual information (MI) according to association rules.

No	Health Association Rules	Lift	MI
1	{Obesity}→{Hypertension}	2.059	1.042
2	{High fat, Chronic Disease}→ {Trans fat, Cancer}	2.333	1.222
3	{Anemia, Diet, Cure}→ {Light head, Vitamin, Fruit}	1.494	0.579
4	{Disk, Pain, Cause}→{Position}	2.059	1.042
5	{Food, Virus}→ {Food Poisoning}	2.059	1.042
6	{Obesity, Complications}→ {Diabetes, Hyperlipidemia}	2.333	1.222
7	{Disk, Stretching}→ {Neck, Waist, Shoulder}	1.707	0.772
8	{Virus}→{Infection}	2.172	1.119
9	{Stress, Management}→ {Exercise, Sleep, Hobby, Relax}	1.537	0.62
10	{Hangover, Cure}→ {Water, Egg, Exercise, Coffee}	2.074	1.052
11	{Cancer, Ablation}→ {Operation, Radiation, Drug}	1.647	0.719
12	{Smoking, Cancer}→ { Laryngeal, Lug, Bladder}	1.537	0.619
13	{Fast-food, Overeating}→ {Cause, Complication, Obesity}	1.463	0.549
....	.....		.....

In Table 3, the Lift value of {Fast-food, Overeating}→ {Cause, Complication, Obesity} is 1.463. When this is substituted into the Equation (2) and calculated, 0.549 is extracted. The Mutual Information value greater than 0 was extracted. This shows that there is a mutual association between {Fast-food, Overeating}→{Cause, Complication, Obesity}. In order to find out how much correlation the mutual information value means, a correlation coefficient is used to identify the degree of correlation [32]. The Cramer C coefficient is used as a correlation coefficient. This is useful for finding the correlation of several items, and the result value is expressed between 0 and 1. In the Cramer C correlation coefficient, the closer the value is to 0, the higher the correlation. The closer the value to 1, the smaller the correlation. Equation (3) shows the Cramer C coefficient.

$$C = \sqrt{\frac{\sum \frac{(O-E)^2}{E}}{N(S-1)}} \quad \begin{array}{l} \text{N : Total number of cases} \\ \text{O : MI(w1,w2)} \\ \text{B : MI(w1,w2)/sum of MI(w1,w2)} \\ \text{S : small number of values in row and columns} \end{array} \tag{3}$$

In equation (3), N represents the total number of cases. O represents the MI of the rule. B represents the value of the mutual information of the rule divided by the sum of the total mutual information. S represents the value of the less item of the rows and columns. For example, in a 10×2 matrix, the value of S is 2. Table 4 is a part of correlation coefficient based on the mutual information.

TABLE 4. Cramer C correlation coefficient according to words.

	Obesity	Infection	Stress	Complications	…
Hypertension	0.045	0.039	0.076	0.133	…
Virus	0.066	0.058	0.033	0.088	…
Relax	0.060	0.091	0.078	0.082	…
Cure	0.070	0.087	0.088	0.092	…
Diet	0.128	0.23	0.067	0.079	…
…	…	…	…	…	…

In Table 4, the mutual information of 10 words of {hypertension, virus, relax, cure, diet, obesity, infection, stress, complications, and management} was constructed as 5×5 matrix to compare the correlation coefficients. For example, the mutual information of {obesity} and {hypertension} is 1.042. Since it was constructed as 5×5 matrix, the value of N means 25, and the value of S means 5. The sum of total mutual information means the sum of 25. When the correlation coefficient is extracted through this, the measured value of 0.045 is extracted. This is closer to 0 than 1, indicating that {obesity} and {hypertension} are highly related.

IV. PERFORMANCE EVALUATION

A. MULTI-LEVEL HEALTH KNOWLEDGE MINING IN P2P EDGE NETWORK

The method proposed in this study was developed using Window10 Pro, Intel (R) Pentium (R) CPU G21020 3.10 GHz, 8 GB RAM, and R Studio 1.1.456. For data, health knowledge for healthcare and disease prevention is collected from P2P edge network in real-time and documents are saved in the text format. Unnecessary elements are removed from the collected health knowledge documents to reduce the size of the data. To improve the accuracy of the analysis, we preprocess the data to generate bag-of-health-words composed of health words. For preprocessing, R's KoNLP package and tm package are used. As a morpheme analysis package for Korean natural language processing, KoNLP extracts nouns from the document. Tm is a package that can be used for text mining. It removes stop words, numbers, punctuation marks, and duplicate words [33]. Using the Apriori algorithm of data mining method and Multi-level association rule, Bag-of-health-word recompose a health transaction in a document unit in order to find the association of words. The association of words is found and the health knowledge tree is created in the health transaction. Also, the knowledge tree and mutual information are used to create health knowledge. Correlation coefficients are used to identify the degree of correlation of

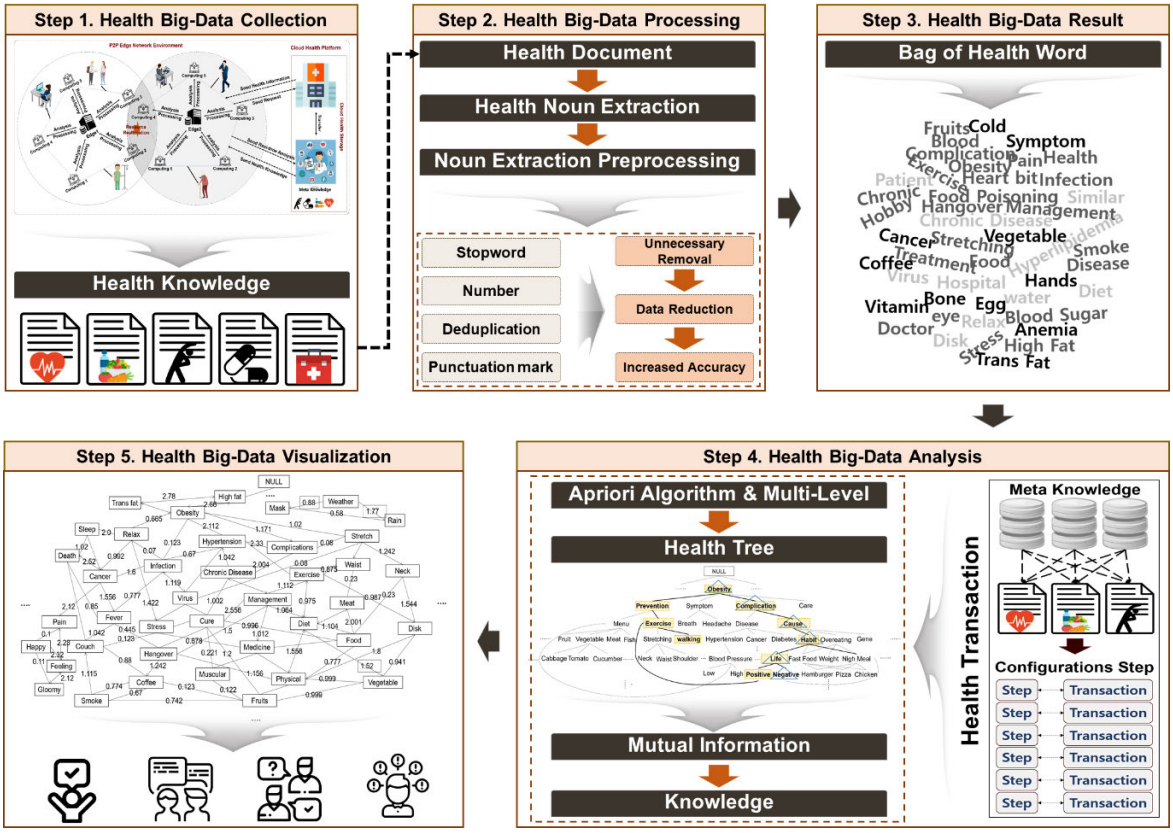

FIGURE 5. Multi-level health knowledge mining process in P2P edge network.

the mutual information. It visualizes the embedding network structure to provide health management information.

The network structure visualizes the association between words and provides users with information so that they can understand it intuitively. The relationships between words can be expressed by a graph consisting only of words and a graph of words appearing at the same time. If we find a hidden relationship such as similarity relationship, antonym relationship, consent relationship, partial relationship between words, it can be represented as a complex network structure [34]. For example, the partial relationships of chronic diseases include obesity, hypertension, diabetes, management, serious, treatment, and recovery. Since management and treatment are the synonym relationship, and recovery and serious are the antonym relationship, they can be expressed as a network structure. Embedding network can express the unique vectors of words as a network structure in a multidimensional space [35]. It can be found that the closer the distance of the words, the higher the correlation. Figure 5 shows the multi-level health knowledge mining process in P2P edge network.

In Figure 5, the user uses smart devices and IoT devices to request health data to health platforms distributed in the edge network. The edge network processes and analyzes the information requested by the user in a distributed health server in real-time and generates and provides meta-information. The meta-information extracts 11,668 nouns through the noun extraction process in the form of text. Through preprocessing, 10,400 words are extracted and construct bag-of-health-words.

Bag-of-health-words construct a transaction for discovery association rules and uses Apriori and multi-level rule algorithms to find the associations. Based on the discovered associations, the degree of association is calculated through the mutual information and use correlation coefficient is used to determine the degree of correlation. This generates the health knowledge and schematizes information into an embedding network structure in order to provide it intuitively.

Figure 6 is the result of Cramer C correlation coefficient based on the mutual information. The y-axis and the x-axis represent correlation coefficients and health words, respectively. In Figure 6, {obesity} has the correlation coefficient of 0.045 with {hypertension}, 0.06 with {virus}, 0.07 with {cure}, 0.066 with {relax} and 0.128 with {diet}. This means that the correlation is high because the correlation coefficient close to 0 was extracted. The word relationship is extended based on the mutual information to visualize it as network structure based on multi-level health knowledge mining process. Figure 7 shows the result of multi-level health knowledge in P2P edge network.

B. PERFORMANCE EVALUATION

The experimental documents for the performance evaluation of the proposed method consist of 1,000 experimental

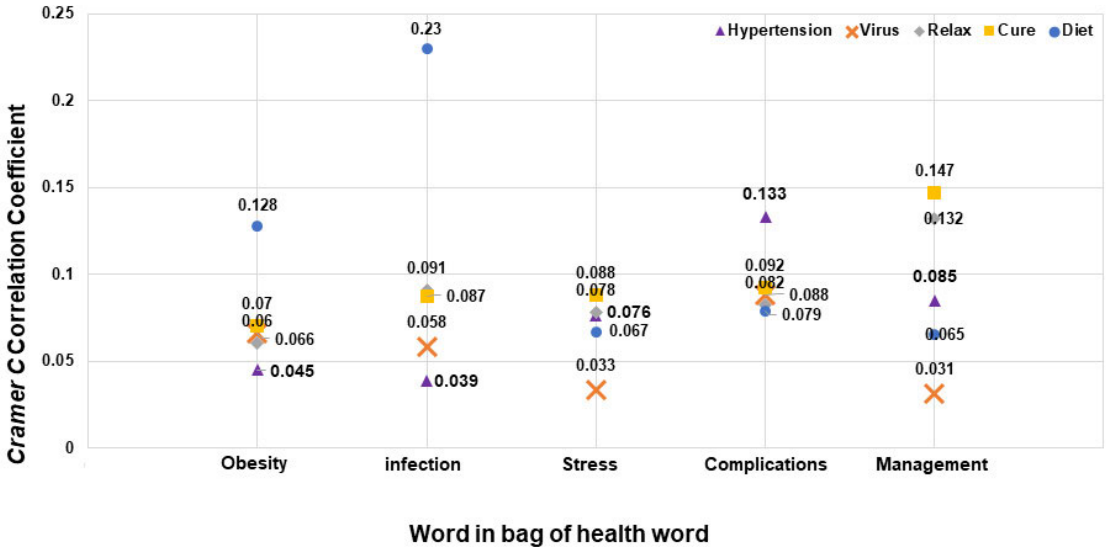

FIGURE 6. Cramer C correlation coefficient according to health words.

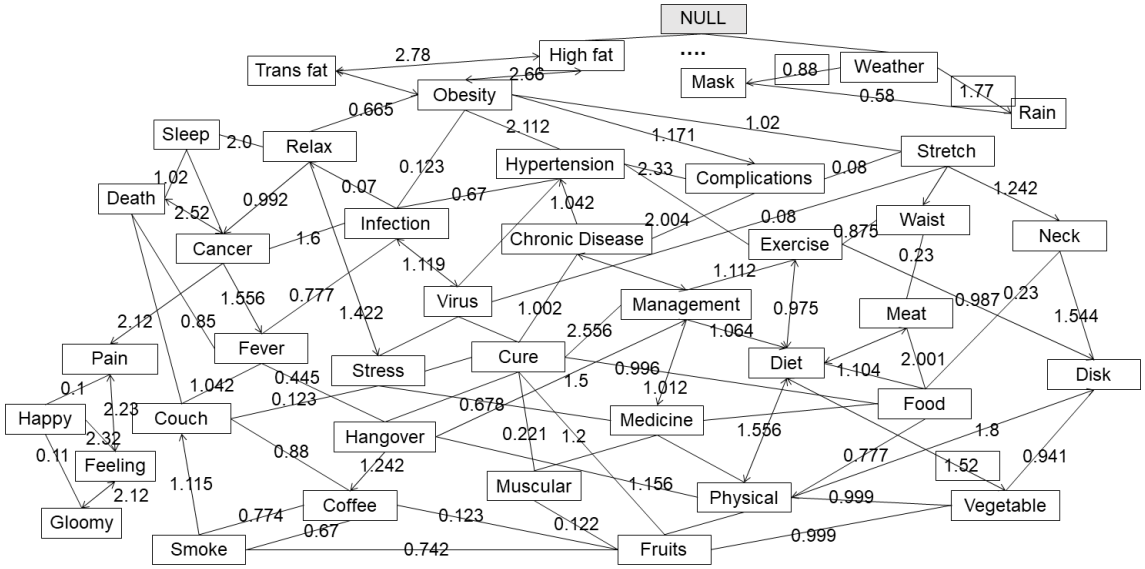

FIGURE 7. Result of multi-level health knowledge in P2P edge network.

documents by combining health information documents collected from P2P edge network and documents provided in the Health Insurance Review and Assessment Service [36].

The experimental documents were composed of 800 training documents and 200 test documents, and this process was repeated 10 times. The health corpus is constructed from the experimental documents through preprocessing. The health corpus is reconstructed into a health transaction and association rules are used according to the proposed method to construct the knowledge base. We evaluate the F-measure using precision and recall according to the change of minimum support in the transaction and whether or not mining is applied. HW is health word used for test, N is words in the whole document and $(HW \cap N)$ is health words in all documents and the number of words in the experimental documents matching with the words used in the test, respectively. Precision refers to the rate of words actually appearing matching with health words in all the documents. Recall refers to the rate of words used in the test matching with words actually appearing. Equation (4) represents the Precision, and Equation (5) represents the recall. In addition, Equation (6) shows an equation for calculating the F-measure using accuracy and recall.

$$\text{Precision} = \frac{(HW \cap N)}{N} \qquad \text{HW : Health words used in the test} \tag{4}$$

$$Recall = \frac{(HW \cap N)}{HW} \qquad \text{N: Words in the whole document} \tag{5}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

The performance evaluation uses the proposed mining-based mutual information (MbMI), existing mining-based word frequency (MbWF) [37], [38], word concurrence frequency (WCoF) [39], [40] in the document to find the relationship between words. It performs performance evaluation while repeatedly changing minimum support. Figure 8 shows the precision and recall according to MbMI, MbWF, and WCoF. The y-axis and the x-axis represent the precision and recall, respectively.

From about 0.4 based on the recall, the MbMI method proposed in Fig. 8 was highly evaluated for recall and accuracy.

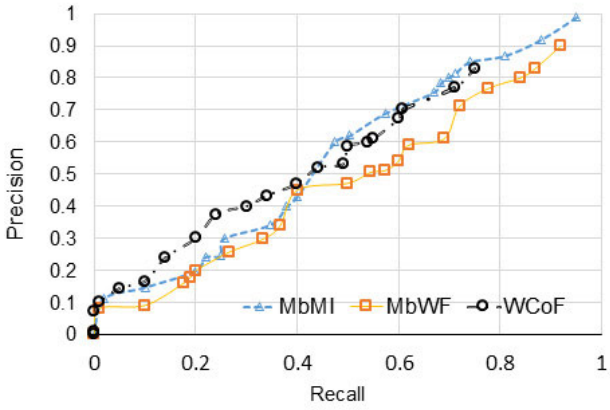

FIGURE 8. Precision and Recall according to MbMI, MbWF, and WCoF.

In the case of WCoF, the number of co-occurrence between words in a document is limited. In addition, recall of association rules was evaluated poorly because the association was not considered. In the case of MbWF, the degree of the mutual association between words was not considered, so a meaningless relationship was found, resulting in poorly evaluated precision. The degree of association using association rules between words and mutual information can improve recall and precision.

Table 5 shows the precision, recall, and F-measure evaluation results of MbMI, MbWF, and WCoF.

TABLE 5. Precision, Recall, and F-measure evaluation results of MbMI, MbWF, and WCoF.

Method	Evaluation	10%	20%	30%	40%	50%
MbMI	Precision	0.870	0.850	0.813	0.803	0.787
	Recall	0.810	0.740	0.711	0.700	0.684
	F-measure	0.839	0.791	0.759	0.748	0.732
MbWF	Precision	0.800	0.767	0.711	0.610	0.590
	Recall	0.840	0.777	0.720	0.690	0.620
	F-measure	0.820	0.772	0.715	0.648	0.605
WCoF	Precision	0.700	0.673	0.610	0.599	0.587
	Recall	0.610	0.60	0.550	0.540	0.500
	F-measure	0.652	0.634	0.578	0.568	0.540

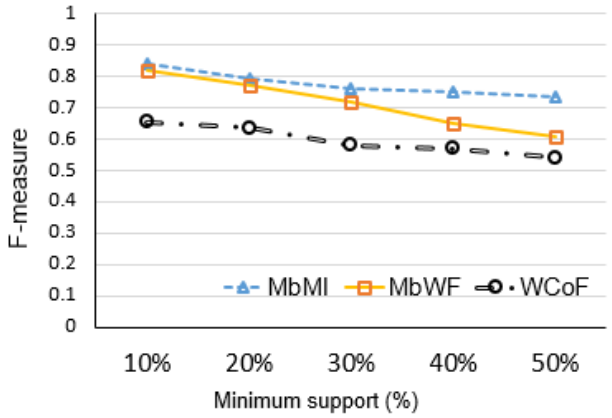

FIGURE 9. F-measure evaluation results.

Figure 9 shows the F-measure evaluation results. The y-axis and the x-axis represent the measured value of the F-measure and the minimum support, respectively. In Fig. 9, the F-measure evaluation results show that the performance of the comparison methods at the minimum support of 10% is excellent. MbMI and MbWF are 0.839 and 0.820 on average, respectively, and WCoF has excellent performance of MbMI method proposed as 0.652 on average. The performance of the proposed MbMI method was found to be excellent when the minimum support is 20%, 30%, 40%, and 50%.

The proposed Lift based mutual information method is evaluated in terms of accuracy. In other words, according to a change in the minimum support, entropy-based information is compared with the proposed improvement based mutual information in terms of accuracy. Fig. 10 shows the result of an accuracy comparison between entropy and Lift based mutual information. The horizontal axis represents the minimum support, and the vertical axis means accuracy.

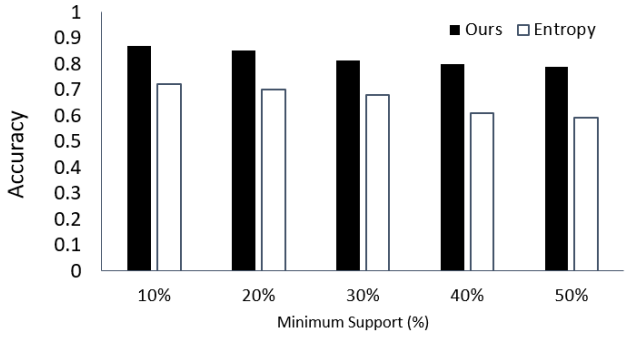

FIGURE 10. The result of an accuracy comparison between entropy and Lift based mutual information.

As shown in Figure 10, the proposed improvement based mutual information had better accuracy. Entropy is the value calculated in the way of multiplying the information of each case by a probability and adding up all. It is ambiguous. In the association rule, the number of cases is changed depending on reliability and support so that it is ambiguous. Therefore, the entropy-based association rule information method has

a limitation. On contrary, the proposed method calculates information on the basis of the association rules extracted, it has better accuracy than the entropy-based method.

For the evaluation of the proposed method, it is compared with conventional association rule-based knowledge mining techniques. In terms of the rule creation according to a support change, F-measure using recall and accuracy is compared. Table 6 shows the result of the performance comparison between the proposed method and the conventional models.

TABLE 6. The result of the performance comparison between the proposed method and the conventional models.

Method	Evaluation	10%	20%	30%	40%	50%
Ours	Precision	0.870	0.850	0.813	0.803	0.787
	Recall	0.810	0.740	0.711	0.700	0.684
	F-measure	0.839	0.791	0.759	0.748	0.732
R. Xu et al. [41]	Precision	0.735	0.705	0.613	0.683	0.658
	Recall	0.713	0.628	0.675	0.713	0.723
	F-measure	0.724	0.665	0.643	0.698	0.689
M. Nasr et al. [42]	Precision	0.682	0.653	0.513	0.588	0.495
	Recall	0.535	0.500	0.651	0.610	0.600
	F-measure	0.600	0.566	0.574	0.599	0.542

As shown in table 6, the proposed method showed the best performance. The method of Xu and Luo [41] generates knowledge through correlation analysis and non-linear modeling. With the use of association rule mining, it finds the relations between risk factors and a manager and does modeling for extracting knowledge through the random-forest algorithm. Since data are limited to dangerous situations, it has low performance. The method of Nasr *et al.* [42] extracts knowledge through the supervised learning technique using data with designated class labels. For this reason, if unlabeled or new data are used for analysis, recall and accuracy are low.

V. CONCLUSION

In this study, we propose a multi-level health knowledge mining process in P2P edge network. The proposed method collects data from P2P networks. The collected data has noise problems such as subjective opinions, distorted information, and exaggerated information. Therefore, precise and reliable information should be provided to users who have difficulty making accurate information decisions. As the number of nodes connected to the P2P network increases, there are an overload problem and a security problem through the Internet network. Overload and security problems are solved by distributed processing of data through P2P edge network. In addition, reliable knowledge should be generated through association and mutual information to solve the noise problems. P2P edge networks can be stored for data reuse and expansion. This allows users to request and collect knowledge necessary for healthcare and disease prevention to the health platform using the smart platform and IoT device anytime and anywhere. The collected data is composed of bag-of-health-words through morpheme analysis and preprocessing. In the health corpus, frequency of word appearance and distribution are used to reconstruct them into a health transaction, and Multi-Level and Apriori Algorithm of Data Mining are used to create a health knowledge tree based on the findings of association rules. The health knowledge tree identifies the various relationships between health words. The mutual information is used based on the health knowledge tree to identify the degree of association and generate health knowledge. In addition, the Cramer C correlation coefficient is used to determine the degree of the mutual association of the MI. Through the Cramer C correlation coefficient, the correlation between several words can be found, which represents a value between 0 and 1. The generated knowledge is visualized and expressed as an embedding network so that the user can intuitively understand it easily. The proposed method in the performance evaluation result was highly evaluated in the F-Measure using precision and recall compared with the existing method. In the P2P edge network, lift-based mutual information can easily identify the meaningful association of words. In addition, the multi-level health knowledge mining process provides meta-knowledge related to healthcare and disease prevention through visualization in real time. In the P2P edge network, users are provided with knowledge services such as health life habit information, disease information (cause, symptom, diagnosis, treatment, prevention) through the association of inquiries and the health knowledge tree.

REFERENCES

[1] H. Yoo and K. Chung, "PHR based diabetes index service model using life behavior analysis," *Wireless Pers. Commun.*, vol. 93, no. 1, pp. 161–174, Mar. 2017.

[2] H. Yoo, S. Han, and K. Chung, "A frequency pattern mining model based on deep neural network for real-time classification of heart conditions," *Healthcare*, vol. 8, no. 3, pp. 234–251, Sep. 2020.

[3] J.-C. Kim and K. Chung, "Multi-modal stacked denoising autoencoder for handling missing data in healthcare big data," *IEEE Access*, vol. 8, pp. 104933–104943, May 2020.

[4] H. Jung and K. Chung, "Knowledge-based dietary nutrition recommendation for obese management," *Inf. Technol. Manage.*, vol. 17, no. 1, pp. 29–42, Mar. 2016.

[5] J.-S. Kang, J.-W. Baek, and K. Chung, "PrefixSpan based pattern mining using time sliding weight from streaming data," *IEEE Access*, vol. 8, pp. 124833–124844, Jul. 2020.

[6] J.-W. Baek and K. Chung, "Context deep neural network model for predicting depression risk using multiple regression," *IEEE Access*, vol. 8, pp. 18171–18181, Jan. 2020.

[7] D.-H. Shin, R. C. Park, and K. Chung, "Decision boundary-based anomaly detection model using improved AnoGAN from ECG data," *IEEE Access*, vol. 8, pp. 108664–108674, Jun. 2020.

[8] *Korea Centers for Disease Control and Prevention*. Accessed: Jun. 19, 2020. [Online]. Available: http://health.cdc.go.kr/

[9] H.-C. Hsieh, C.-S. Lee, and J.-L. Chen, "Mobile edge computing platform with container-based virtualization technology for IoT applications," *Wireless Pers. Commun.*, vol. 102, no. 1, pp. 527–542, May 2018.

[10] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Comput. Netw.*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.

[11] K. Chung, H. Yoo, and D. E. Choe, "Ambient context-based modeling for health risk assessment using deep neural network," *J. Ambient Intell. Hum. Comput.*, vol. 11, no. 4, pp. 1387–1395, Apr. 2020.
[12] S.-H. Kim and K. Chung, "Emergency situation monitoring service using context motion tracking of chronic disease patients," *Cluster Comput.*, vol. 18, no. 2, pp. 747–759, Jun. 2015.
[13] Samsung Electronics. *Samsung Health*. Accessed: Jun. 19, 2020. [Online]. Available: https://www.samsung.com/
[14] *InBodyBand*. Accessed: Jun. 19, 2020. [Online]. Available: http://www.inbody.co.kr/
[15] P. Corcoran and S. K. Datta, "Mobile-edge computing and the Internet of Things for consumers: Extending cloud computing and services to the edge of the network," *IEEE Consum. Electron. Mag.*, vol. 5, no. 4, pp. 73–74, Oct. 2016.
[16] E.-Y. Jung, J.-H. Kim, K.-Y. Chung, and D. K. Park, "Home health gateway based healthcare services through U-Health platform," *Wireless Pers. Commun.*, vol. 73, no. 2, pp. 207–218, Jun. 2013.
[17] J.-C. Kim and K. Chung, "Mining health-risk factors using PHR similarity in a hybrid P2P network," *Peer Peer Netw. Appl.*, vol. 11, no. 6, pp. 1278–1287, Feb. 2018.
[18] K. Chung and R. C. Park, "P2P cloud network services for IoT based disaster situations information," *Peer Peer Netw. Appl.*, vol. 9, no. 3, pp. 566–577, May 2016.
[19] M. Chen, W. Li, Y. Hao, Y. Qian, and I. Humar, "Edge cognitive computing based smart healthcare system," *Future Gener. Comput. Syst.*, vol. 86, pp. 403–411, Sep. 2018.
[20] J. Li, J. Wu, and L. Chen, "Block-secure: Blockchain based scheme for secure P2P cloud storage," *Inf. Sci.*, vol. 465, pp. 219–231, Oct. 2018.
[21] S. Jabbour, F. E. El Mazouri, and L. Sais, "Mining negatives association rules using constraints," *Procedia Comput. Sci.*, vol. 127, pp. 481–488, Jan. 2018.
[22] D.-H. Shin, M.-J. Kim, S. Oh, and K. Chung, "Knowledge reasoning model using association rules and clustering analysis of multi-context," *J. Korea Converg. Soc.*, vol. 10, no. 9, pp. 11–16, 2019.
[23] A. Bhandari, A. Gupta, and D. Das, "Improvised apriori algorithm using frequent pattern tree for real time applications in data mining," *Procedia Comput. Sci.*, vol. 46, pp. 644–651, Nov. 2015.
[24] H. J. Kim, J. W. Baek, and K. Chung, "Optimization of associative knowledge graph using TF-IDF based ranking score," *Appl. Sci.*, vol. 10, no. 13, pp. 4590–4610, Jan. 2020.
[25] H. Jung, H. Yoo, and K. Chung, "Associative context mining for ontology-driven hidden knowledge discovery," *Cluster Comput.*, vol. 19, no. 4, pp. 2261–2271, Dec. 2016.
[26] J.-C. Kim and K. Chung, "Associative feature information extraction using text mining from health big data," *Wireless Pers. Commun.*, vol. 105, no. 2, pp. 691–707, Mar. 2019.
[27] I. Fortin and D. Oliver, "To imitate or differentiate: Cross-level identity work in an innovation network," *Scand. J. Manage.*, vol. 32, no. 4, pp. 197–208, Dec. 2016.
[28] I. Kim, M. Springer, Z. G. Zhang, and Y.-S. Park, "Organizational learning: Approximation of multiple-level learning and forgetting by an aggregated single-level model," *Comput. Ind. Eng.*, vol. 131, pp. 442–454, May 2019.
[29] E. Baralis, L. Cagliero, T. Cerquitelli, and P. Garza, "Generalized association rule mining with constraints," *Inf. Sci.*, vol. 194, pp. 68–84, Jul. 2012.
[30] K. Dong, L. Long, H. Zhang, and Y. Gao, "The mutual information based minimum spanning tree to detect and evaluate dependencies between aero-engine gas path system variables," *Phys. A, Stat. Mech. Appl.*, vol. 506, pp. 248–253, Sep. 2018.
[31] S. A. Peixoto and P. P. R. Filho, "Neurologist-level classification of stroke using a structural co-occurrence matrix based on the frequency domain," *Comput. Electr. Eng.*, vol. 71, pp. 398–407, Oct. 2018.
[32] H. Akoglu, "User's guide to correlation coefficients," *Turkish J. Emergency Med.*, vol. 18, no. 3, pp. 91–93, Sep. 2018.
[33] *The R Foundation*. Accessed: Jun. 27, 2020. [Online]. Available: https://cran.r-project.org/
[34] M. Kudelka, P. Kromer, M. Radvansky, Z. Horak, and V. Snasel, "Efficient visualization of social networks based on modified Sammon's mapping," *Swarm Evol. Comput.*, vol. 25, pp. 63–71, Dec. 2015.
[35] J. Chen, Y. Tao, and H. Lin, "Visual exploration and comparison of word embeddings," *J. Vis. Lang. Comput.*, vol. 48, pp. 178–186, Oct. 2018.
[36] *Health Insurance Review and Assessment Service*. Accessed: Jun. 27, 2020. [Online]. Available: http://www.hira.or.kr/
[37] B. C. Hidayanto, R. F. Muhammad, R. P. Kusumawardani, and A. Syafaat, "Network intrusion detection systems analysis using frequent item set mining algorithm FP-max and Apriori," *Procedia Comput. Sci.*, vol. 124, pp. 751–758, Jan. 2017.
[38] A. Govada, A. Patluri, and A. Honnalgere, "Association rule mining using Apriori for large and growing datasets under Hadoop," in *Proc. Int. Conf. Netw., Commun. Comput.*, Dec. 2017, pp. 14–17.
[39] D. Paperno, M. Marelli, K. Tentori, and M. Baroni, "Corpus-based estimates of word association predict biases in judgment of word co-occurrence likelihood," *Cognit. Psychol.*, vol. 74, pp. 66–83, Nov. 2014.
[40] P. Agustí, V. J. Traver, and F. Pla, "Bag-of-words with aggregated temporal pair-wise word co-occurrence for human action recognition," *Pattern Recognit. Lett.*, vol. 49, pp. 224–230, Nov. 2014.
[41] R. Xu and F. Luo, "Risk prediction and early warning for air traffic controllers' unsafe acts using association rule mining and random forest," *Saf. Sci.*, vol. 135, pp. 105125–105134, Mar. 2021.
[42] M. Nasr, M. Hamdy, D. Hegazy, and K. Bahnasy, "An efficient algorithm for unique class association rule mining," *Expert Syst. Appl.*, vol. 164, pp. 113978–113987, Feb. 2021.

JI-WON BAEK received the B.S. degree from the School of Computer Information Engineering, Sangji University, South Korea, in 2017, and the M.S. degree from the Department of Computer Science, Kyonggi University, South Korea, in 2020, where she is currently pursuing the Ph.D. degree with the Department of Computer Science. She worked for Data Management Department, Infiniq Company, Ltd. She has been a Researcher with the Data Mining Laboratory, Kyonggi University. Her research interests include data mining, data management, knowledge systems, automotive testing, deep learning, healthcare, and recommendation.

KYUNGYONG CHUNG received the B.S., M.S., and Ph.D. degrees from the Department of Computer Information Engineering, Inha University, South Korea, in 2000, 2002, and 2005, respectively. He has worked for the Software Technology Leading Department, Korea IT Industry Promotion Agency (KIPA). From 2006 to 2016, he was a Professor with the School of Computer Information Engineering, Sangji University, South Korea. Since 2017, he has been a Professor with the Division of AI Computer Science and Engineering, Kyonggi University, South Korea. He was named a 2017 Highly Cited Researcher by Clarivate Analytics. His research interests include data mining, artificial intelligence, healthcare, knowledge systems, HCI, and recommendation systems.