Application of Data Mining Techniques in a Personalized Diet Recommendation System for Cancer Patients

[1]Wahidah Husain, [2]Lee Jing Wei, [3]Sooi Li Cheng and [4]Nasriah Zakaria
School of Computer Science,
Universiti Sains Malaysia,
11800, USM,
Penang, Malaysia.
[1]wahidah@cs.usm.my, [2]leejingwee@yahoo.com, [3]lichengsooi@yahoo.com, [4]nasriah@cs.usm.my

Abstract— **Cancer is one of the top fatal diseases in the world. When cancer patients undergo treatment, they need a different diet to help them withstand the side-effects of treatment, as well as to provide sufficient nutrition to boost the recovery cycle. Most of the dietician systems currently in the market are more in the advice area rather than recommending specific diet menus for users. These systems usually provide advice in a general form, for example: "Cancer patients should eat foods with high protein". Such recommendations are insufficient to assist cancer patients with the physical preparation necessary to withstand the side effects of treatment, or with ensuring they take adequate nutrients for their body. We propose a Personalized Diet Recommendation System for Cancer Patients to help patients manage their daily food intake. The proposed system integrates the data mining techniques of Case-based Reasoning, Rule-based Reasoning and Genetic Algorithm. Case-based Reasoning is used to suggest a set of diet plans taken from the cases existing in the system, whereas Rule-based Reasoning is used to filter out irrelevant cases from the system and select the most appropriate case to be suggested to the patient. The Genetic Algorithm technique ensures that the diet menus suggested are customized according to each patient's personal health conditions. The output of the diet plan system is in the form of a list of specific nutritional values to be taken daily, and a menu recommendation suggesting actual dishes for the patient.**

Keywords- cancer patients; genetic algorithm; rule-based reasoning; case-based reasoning; data mining; diet recommendations

I. INTRODUCTION

According to the Malaysia National Cancer Registry, there were a total of 67,792 cancer patients in Malaysia between 2003 and 2005 [1]. There is approximately 1 cancer patient per 7 men and per 6 women. The data also revealed that lung and nasopharyngeal cancers are common among males, whereas breast and cervical cancers are common among females.

References show that dietary factors play a role of about 30% in cancer occurrence -- that is, an unhealthy daily diet raises the risk of cancer for human beings [2]. For example, excessive intake of saturated fat may heighten the risk of breast, uterus, ovary and colon cancer, while a serious lack of fiber intake can raise the risk of colon cancer. Diet not only plays an important role in causing cancers, it also has a significant effect on patient recovery. However, there is currently no system to help cancer patients manage their daily diet, as all existing systems provide diet information in a general format. Incorporation of existing mature computing power into diet system development can maximize the benefits to cancer patients in terms of creating and preserving a healthier living style and an effective recovery cycle.

The Personalized Diet Recommendation System is designed to deal with daily diet planning for cancer patients and to construct diet menus corresponding to their needs from the diet plans available. The system incorporates Case-based Reasoning, Rule-based Reasoning and Genetic Algorithm in the design of modules related to diet planning and menu construction. The proposed system will provide advice to the patient in the form of total nutrition components to be taken daily, as well as suggested dishes for the type of diet menu corresponding to the total nutrition advised. Cancer patients could thus obtain maximum assistance in respect to their ideal food intake. All result sets provided from the system are customized personally to each patient's personal health information, eating habits and food preferences.

The main purpose of the system is to provide assistance to users of the system (cancer patients and patient care takers) in the area of daily diet. To achieve this purpose, the system focuses on calculating the daily diet required by a patient in terms of nutrition values corresponding to their personal health condition. Consequently, the system proposes for each patient a daily diet menu with a list of recommended dishes for breakfast, lunch and dinner.

II. BACKGROUND STUDY

A. Cancer

Cancer, also known as "malignancy", refers to a class of diseases which exhibit uncontrollable abnormal growth in certain groups of body cells. These growths are able to infest other body t [2]issues and spread to other organs through the blood and lymph systems [3]. There are more than 100 types of cancers; the most well-known are cancers of the breast, skin, lung, colon, prostate, and lymphatic system. Symptoms of

978-1-4673-0020-9/11/$26.00 ©2011 IEEE

cancer vary widely depending on the type of the cancer. The medical community has defined 5 stages of cancer, starting with stage 0 as minor level and stage 4, in which the cancer has begun to metastasize (spread elsewhere in the body).

Common cancer treatments include chemotherapy, radiotherapy, hormone therapy and surgery [3]. Chemotherapy uses drugs to kill fast-growing cancer cells. It is often used after surgery to lower the risk of the cancer recurring. Radiotherapy, also known as radiation therapy, x-ray therapy, or irradiation, uses a certain type of energy called ionizing radiation to kill cancer cells and shrink tumors [3]. However, it injures or destroys all cells in the treated area by damaging their genetic material, making it impossible for the cells to undergo normal mutation and growth [2]. The goal of radiotherapy is to damage as many cancer cells as possible, so as to limit its harm to nearby healthy tissue. The different treatments have different side-effects, including diarrhea, constipation and nausea. These side-effects obviously affect the patient's appetite and eating habits. Thus, it is important to carefully monitor the patient's food intake to ensure the maximum recovery effect [4].

B. Role of Food in Cancer Patients' Recovery Cycle

Diet plays an important role in determining how successful a treatment is for a cancer patient. Eating the right kinds of foods before, during and after treatment can boost the effect of the treatment and help the patients stay stronger. To ensure proper nutrition is absorbed, cancer patients require foods that contain key nutrients like vitamins, minerals, protein, carbohydrates, fat and water [6].

However, the side effects of treatment often cause patients to face appetite problems, making them unable to eat well. Symptoms that interfere with the eating habits of cancer patients include anorexia, nausea, vomiting, mouth sores, trouble swallowing, pain and anxiety [7]. Patients will likely suffer loss of appetite and may lose the ability to smell and taste, which decreases the ability to eat and absorb nutrition from foods. Thus, patients are likely to gain insufficient nutrition to prepare their bodies for effective treatment, causing them to be weak, tired and unable to resist infections or withstand the necessary therapy. The ability of the body to fight off the side effects of the treatment will decrease tremendously, and could cause death in extreme cases [7]. From this, we can see that food plays a decisive role in the recovery cycle for cancer patients.

C. Case-based Reasoning

Case-based Reasoning (CBR) is a data mining technique for problem solving. CBR utilizes the knowledge of previously-experienced concrete problem situations, known as cases, instead of relying solely on general knowledge of the problem domain, or making associations or generalized relationships between descriptors and conclusions [8]. In CBR, a new problem is solved by searching through similar past cases, finding the closest match, and reusing it in the new problem. Any new solutions are retained as new cases and are immediately available for use in solving future problems. In short, CBR is a cyclic and integrated process of solving a problem by applying the experience from past cases to new problems.

CBR possesses advantages and disadvantages. Its advantages are that it proposes solutions to problems quickly, evaluates the solutions when no algorithmic method is available, helps to stay focused on the major part of the problem and alert to previous experience in solving new and existing problems [9]. On the other hand, its disadvantages are that it tends to apply old cases as solutions blindly without validating them in new situations, and that the cases utilized for new solutions may be too biased (i.e., not a close enough match) to solve the new problem. In short, the effectiveness and efficiency of CBR increase in proportion to the number of cases available in the case base.

D. Rule-based Reasoning

Rule-based Reasoning (RBR), another data mining technique, requires elicitation of an explicit model of the problem domain [11]. RBR attempts to capture the extensive knowledge of domain experts and translate it into expressions, so that it can be stated using "what-if" or "if-then-else" analysis. Compilation of the rules then enables a given problem to be evaluated against the rule-base by chaining rules together until a conclusion is reached.

Implementation of RBR in system development eases the process of storing large amounts of information in the case-base, since clarification of rules will help the extraction of the information process [12]. This makes the system more effective in processing the queries submitted by the users. However, sometime, it is difficult to transfer the expert knowledge into distinct rules.

In terms of information utilization, RBR adds a filtering feature, since only information that fulfills the rules applied is retrieved for utilization. Since evaluation of rules is done from time to time in the process of retrieval, the scope of information retrieval is narrowed down quickly. This makes the system function in a more effective and efficient manner, since irrelevant information is omitted from consideration of the system for further processing [11].

E. Genetic Algorithm

Genetic Algorithm (GA) is an adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics [13]. It is designed to simulate the process of evolution, to fulfill requirements in an evolutionary manner by selecting according to the requirements. GA systems are modeled loosely on the principles of evolution via natural selection through variation-inducing operators such as mutation and crossover [14]. An effective GA representation and meaningful fitness evaluation is the key to success [15]. GA is capable of discovering good solutions rapidly for difficult and multi-dimensional problems; it is most useful and efficient when the search space is large, complex or poorly understood, and when the domain knowledge is limited or the expert knowledge is difficult to encode in rules.

The advantage of the GA approach is the ease with which it can handle arbitrary kinds of constraints and objectives, and that it provides workable solutions that meet, or closely approach, the requirements of a given situation [13]. Though it might not find the best solution, more often than not it comes up with a partially optimal solution due to the repetitive mutation, crossover, inversion and selection operators embedded within the GA approach.

F. Existing Diet Systems

Most existing diet recommendation systems incorporate CBR or RBR, or combine these two techniques [9] into a hybrid CBR-RBR system. Examples include Case-Base Menu Planner (CAMP), CAMP Enhanced by Rules (CAMPER) and Pattern Regulator for the Intelligent Selection of Menus (PRISM) [10].

CAMP's case base holds 84 daily menus, all of which fulfill the Reference Daily Intake (RDI) as determined by the Food and Drug Administration, and the Dietary Guidelines for Americans as laid out by the Center for Nutrition Policy and Promotion. Cases in CAMP contain past solutions and features that indicate when the solution is likely to be useful again. CAMP operates by retrieving and adapting daily menus from its case base [16]. Reusability metric is used to select and retrieve cases based on the ease of adapting it to current goals. Comparison is done against all constraints to find the best case out of the available cases, as there is no one menu that can fill all different requirements of the users. The case that is the closest match to the current problem is selected, and then adapted until it meets all the constraints specified by the user and the RDI level.

PRISM is a rule-based menu planner that can plan menus for a wide range of dietary requirements [16]. PRISM relies on menu and meal patterns to generate menus which match the required nutrient constraints. A typical menu will contain breakfast, lunch, dinner and snack. PRISM uses four databases (meal types, dish types, food types and foods) to implement a multilayered hierarchical structure. This multilayered network implements a context-free grammar for the production of menus. Nutrition constraints cannot be built into menus up front because they are not context-free, so PRISM uses backtracking to alter the menu. When the original menu comes close to meeting constraints, alteration can produce a menu that meets all constraints by substituting new dishes for those in the diet menu that were found nutritionally lacking. Solving one problem, however, sometimes creates more problems. After the menu is generated, it is shown to the user. The user can choose to delete or add foods to the menu. PRISM records the effects on the overall nutritional value of the menu, and can evaluate the trade-offs when the user chooses to change any of the foods in the menu.

CAMPER is a combination of CAMP and PRISM [17]. As described above, CAMP is able to satisfy multiple nutrition constraints and PRISM can construct diet menus in a wide variety of ways using RBR. By merging CAMP's CBR and PRISM's RBR approaches into the CAMPER system, the system not only expands its database but also can construct diet menus in various ways. The CBR and RBR approaches in CAMPER are complementary to each other: CBR contributes an initial menu that meets the constraints by building on food combinations that have proven satisfactory in the past, while RBR allows the analysis of alternatives, so that innovation become possible and thus creates diversity in the diet menus produced by the system. This ability to produce new diet menus improves the capability and the performance of the system over time, as it increases the knowledge database of the system. CAMPER is therefore superior to either CAMP or PRISM alone; as it offers features which neither CAMP nor PRISM could provide [15]. The disadvantage of CAMPER is that it only derives prototypes from specific cases by using the existing menu patterns. It does not employ induction in any formal sense [17].

III. PROPOSED SYSTEM ARCHITECTURE

Fig. 1 shows the four components of the proposed system architecture: user management module, diet planning module, menu construction module and menu adaptation module.

A. User management Module

The user management module is responsible for handling all operations related to users (patients/care takers), such as login to the system and the registration of new user accounts.

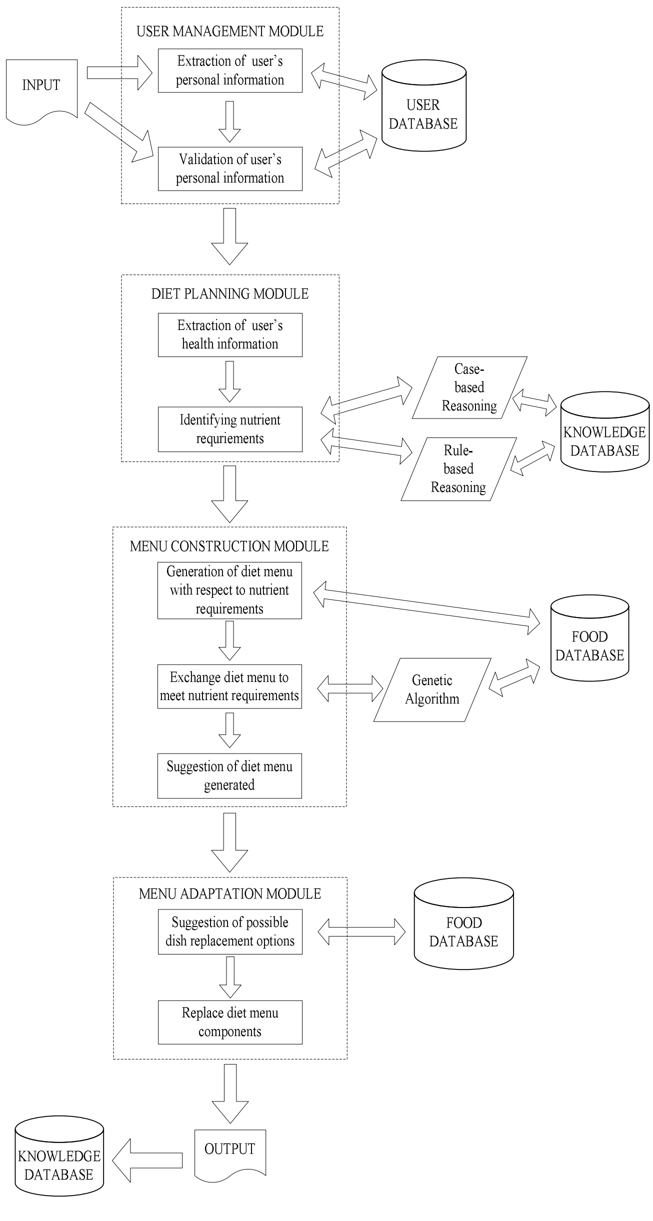

Figure 1: Proposed System Architecture

Information about each user, such as personal medical information or food preferences, is handled under this component as well.

B. Diet Planning Module

In this module, CBR and RBR are applied together with a reusability matrix as shown in Fig. 2. Existing cases in the database serve as the expert knowledge for the computation of this module. First, user information is collected, for subsequent use in CBR. CBR is done in parallel with RBR, so that unnecessary and irrelevant cases are filtered out during retrieval, thus increasing the performance of the CBR. Narrowing of scope is executed if no existing case can fulfill all the conditions submitted, and the results collected are further processed.

Next, the reusability matrix is applied. This concept was chosen so that selected cases from the database will be sure to satisfy the maximum requirements of the user. It preserves the accuracy of the overall computation of the system, as only the closest existing case is chosen for further processing. This is done using a penalty score assignment approach, in which scores are assigned to all the cases to evaluate the degree of bias corresponding to the conditions to be fulfilled. Only the case with the lowest penalty score (meaning that it fulfills more of the requirements than any other case) is selected. Its data is retrieved for further processing and subsequently utilized as the input for the next module of the system, Menu Construction.

C. Menu Construction Module

For the Menu Construction Module, the fundamental concepts of Genetic Algorithm, including selection, crossover and mutation, are applied [5]. The process flow of this module is illustrated in Fig 3. First, a dish is selected from the database according to the user's health condition and preferences, which eliminates the possibility of selecting a dish which is prohibited to the user (rule-based selection*)*. This approach is likely to increase the level of acceptability of the final diet menu, as foods which are prohibited or disliked will not be included.

Crossover is applied during the generation of the diet menus, according to the nutrition values of the Diet Planning Module. A number of unique diet menus, for example 50, are generated. Each of the menus goes through a nutritional values computation process in order to calculate and identify the one that is the best match.

Mutation is applied to ensure that the diet menu completely fulfills the nutrition values calculated by the Diet Planning Module. If the diet menu falls short of (or exceeds) the recommended nutrition values, one of the dishes in the diet menu is replaced with another that has higher (or lower) nutrition values, in order to reduce the gap between the proposed diet menu and the nutrition values calculated from Diet Planning Module. This step continues until the calculated nutrition values of the menu as a whole achieve the desired levels.

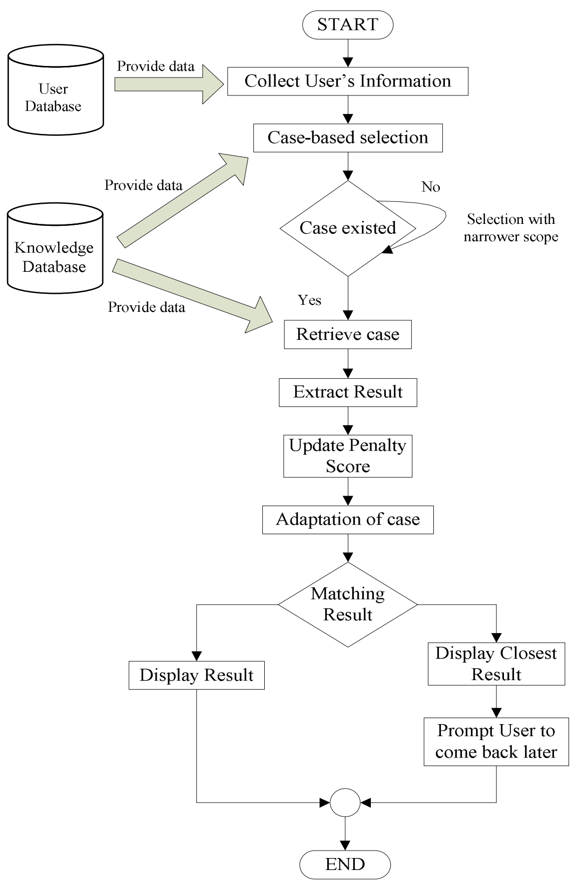

Figure 2. Process Flow of Diet Planning Module

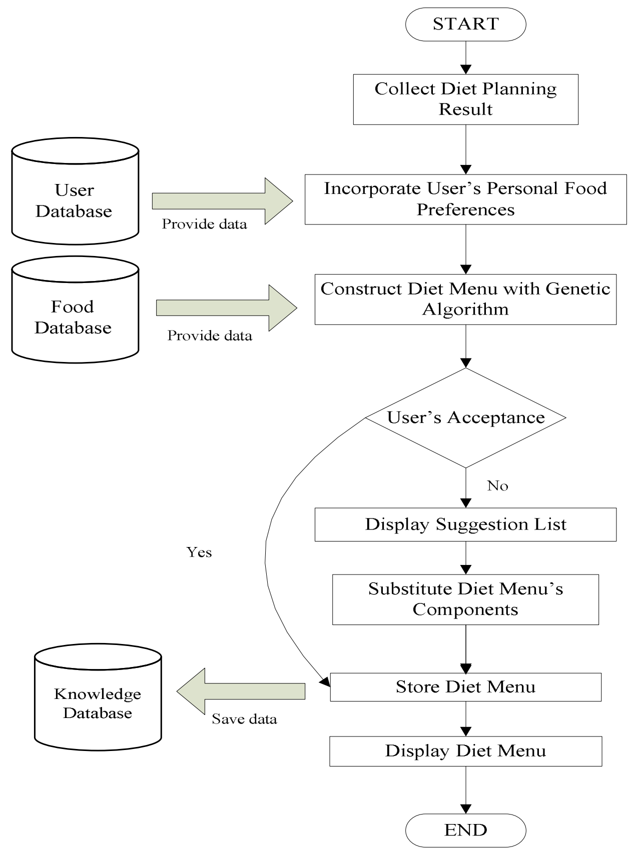

Figure 3. Process Flow of Menu Construction Module

After the best-matching diet menu is created, the result is displayed to the user, who may choose to accept or adapt (modify) the diet menu. The diet menu is inserted into the database only after the user chooses to accept it; otherwise it is stored in a temporary table for later reuse, to ease the process of dish computation and adaptation.

D. Menu Adaptation Module

This function is a subdivision of the Menu Construction Module, and deals with providing suggestions to users wishing to substitute components in the diet menu. To adapt the diet menu, the user first chooses the particular dish he wishes to adapt. Rule-based reasoning (RBR) is applied in this module, so that the system selects from the food database and displays to the user only dishes that are of a type similar to that of the component chosen for substitution. Implementation of RBR indirectly speeds up the system's selection of suggested dishes for display. Thus, no irrelevant dishes are displayed as alternatives for the user; all substitutes or options that are displayed in the suggestion list are certain to correspond to the dish to be replaced and to the user's personal health condition.

IV. IMPLEMENTATION

A bottom-up approach is applied during system development; this eases the quality management process of the system development as we can conduct testing of each of the components individually from time to time, before integrating them to function as one system. It also simplifies the process of handling changes in the requirements and design of the system, as evaluation can be conducted for each component along the development lifecycle.

In order to satisfy diverse requirements, the knowledge base needs to contain a large number of existing cases, so that the probability of generating compatible result sets that meet the user's requirements is increased. On the other hand, RBR itself can only enhance the speed in searching through existing cases inside the knowledge base, as it mainly provides a filtering function for the testing of existing solutions. All irrelevant cases are filtered out during the search process, yet it cannot always find an existing case corresponding to 100% of the user's requirements. The use of Genetic Algorithm can ensure the availability of unique result sets, as it can modify and change existing result sets to generate new ones that can accommodate and fulfill all the requirements of the user.

V. RESULTS AND DISCUSSION

In order to demonstrate the usefulness and accuracy of the system, a few set of inputs are used as shown in Table 1. The output generated by the system based on the input provided by the user is shown in Tables 2 and 3. For example, since this user choose “No Red Meat” in her food preferences and is allergic to eggs, any menu generated for her will not contain any red meat or egg dishes. The nutrition values are also traced to ensure that it fulfills the requirements.

The diet menu generated by the system may be "biased," that is it may not precisely match the recommended nutrition values, as shown in Table 4. Comparing this to Table 2 above, we can see that the nutrition value obtained from the system contains bias in term of certain nutrition values. Such a condition arises due to the limited dish data stored in the database. The algorithm can only meet the nutrition values to the extent that there is sufficient data to be manipulated.

TABLE 1 PERSONAL AND MEDICAL PROFILE

Age	23
Gender	Female
Cancer	Breast Cancer
Stage of Cancer	Stage 2
Treatment	Surgery
Stage of Treatment	After Treatment
Level of Activity	Moderate
Food Preference	No red meat
Allergy	Egg
Ethnic	Malay
Side Effect	Constipation

TABLE 2. DIET PLANNING RESULT

Energy	2567
Carbohydrate	433
Protein	242.1
Calcium	174.3
Thiamin	20
Niacin	13.7
Riboflavin	24.2
Vitamin A	12.1
Vitamin C	3.6
Vitamin D	7.4
Vitamin E	5.2

TABLE 3. DIET MENU RESULT

Daily Diet Menu
Breakfast
1 cup of Milo
2 pieces of Bread
Lunch
1 plate of White Rice
1 bowl of Chicken Soup
1 plate of steamed broccoli
1 plate of Dried Dozen Tofu
Dinner
1 plate of Cereal Rice
1 plate of Tuna fish salad
1 plate of Bread Sticks
1 bowl of Tomato Soup

TABLE 4. NUTRITION VALUE OF DIET MENU RESULT

Energy	2214
Carbohydrate	389
Protein	253.1
Calcium	164.2
Thiamin	13
Niacin	10.1
Riboflavin	13.2
Vitamin A	4.2
Vitamin C	2.7
Vitamin D	5.1
Vitamin E	4.2

Limited data in the database narrows the scope of possibilities for selecting dishes to fulfill 100% of the nutrition value, corresponding to the Diet Planning Result. Thus, in order to ensure maximum fulfillment from the system, the dish database should be as large as possible, so that the algorithm can easily find a set of diet menus containing more than enough dishes to select in order to fulfill the nutritional requirements of the Diet Planning Result.

VI. CONCLUSION AND FUTURE WORK

The application of data mining techniques in a dietician system is highly feasible since they can handle complex processing with high accuracy. However, one limitation of the Personalized Diet Recommendation System for Cancer Patients as described in this paper is that it could only generate diet menu on a daily basis. This is due to the limited number of data about dishes available in the database, which is insufficient for generating a week's worth of menus with unique dishes that achieve weekly nutritional requirements. Thus, the database of dishes should be enriched as much as possible, so that users could generate a weekly diet menu; this will simplify shopping, preparation and planning. Besides that, the database should be enhanced with more complete nutritional information about each dish, so that the algorithm coverage could be wider, preserving a higher accuracy result from the system.

ACKNOWLEDGMENT

We would like to express our gratitude to Universiti Sains Malaysia, Penang for supporting this research.

REFERENCES

[1] http://www.acrm.org.my/ncpr, "Malaysia National Cancer Registry ", [Accessed: 10th October, 2011].

[2] http://www.webmd.com, "WebMD – Better Information, Better Health", [Accessed: 24th April 2011].

[3] http://www.cancer.gov, "National Cancer Institute, US National Institute of Health", [Accessed: 25th April 2011].

[4] http://www.livestrong.com, "3 Non-Nutrient Compounds that Play an Important Role in the Prevention of Cancer", [Assessed: 26th April 2011].

[5] http://www.health360.info, "The Role of Food and Nutrition in Cancer by Dr. Chris Jones", [Accessed: 26th April 2011].

[6] http://www.news-medical.net, "The Medical News", [Accessed: 27th April 2011].

[7] http://www.naturalhealthcourses.com, "What Role Does Food Play in Our Lives?", Elizabeth Springer-Bawauh, [Accessed: 29th April 2011].

[8] A. Aamodt, E. Plaza, "Case-based Reasoning, Foundation Issues, Methodological Variations and System Approaches", AICom–Artificial Intelligence Communication, IOS Press, Vol.7, No. 1, pp 39-59, 1994.

[9] M.M. Richter, A. Aamodt, "Case-based Reasoning Foundation", Cambridge University Press: The Knowledge Engineering Review, Vol.20, No. 3 , pp. 203-207, 2006.

[10] O.P. Rishi, R.Govil and M.Sinha, "Distributed Case-based Reasoning for Intelligent Tutoring System: An Agent Based Student Modeling Diagram", World Academy of Science, Engineering and Technology 29, 2007.

[11] K. Pal, J. A. Campbell, "An Application of Rule-based and Case-based Reasoning within a Single Legal Knowledge Based System", The database for Advances in Information Systems, Vol.28, No. 4, pp 48-63, 1997.

[12] G. L. Fisher, D. E. Busse, "Adding Rule-based Reasoning to Demonstrational Interface Builder", DAWolber, Proceedings of 5th Annual Symposium on User Interface Software and Technology, pp 89-97, 1992.

[13] R. H. Chi, M. Y. Kiang, "An Integrated Approach of Rule-based and Case-based Reasoning for Decision Support", Proceedings of 19th Annual Conference on Computer Science and ACM Annual Computer Science Conference, pp 255-267, 1991.

[14] [http://pesona.mmu.edu.my/~ytbau/tes3211/expert_systems_clips.html, "CLIPS (C Language Integrated Production System) Programming Cheat Sheet", [Accessed: 15th April 2011].

[15] O. J. Mengshoel, D. E. Goldberg, "The Crowding Approach to Niching in Genetic Algorithms", Journal of Evolutionary Computation, Vol.16 Issue 3, pp 315-354, 2008.

[16] C. R. Marling, G. J. Petot, L. S. Sterling, "Integrating Case-based and Rule-based Reasoning to Meet Multiple Design Constraints", Computational Intelligence, Vol.15, No.3, Year 1999.