2

Knowledge Management in Bio-Information Systems

Kuodi Jian
Metropolitan State University, Saint Paul, MN
USA

1. Introduction

Knowledge management is a broad topic. For different people, it may mean different things. For business people, this phrase means the accumulated procedures/processes and experiences (organizational assets), and the way to facilitate the use, and to retain these assets within an organization. The Wikipedia website has the following definition for the knowledge management:

"Knowledge Management (KM) comprises a range of strategies and practices used in an organization to identify, create, represent, distribute, and enable adoption of insights and experiences. Such insights and experiences comprise knowledge, either embodied in individuals or embedded in organizational processes or practice." (Internet resource: http://en.wikipedia.org/wiki/Knowledge_management. Retrieved on 5/30/2011)

For computer science people, especially for those expert system developers, the term "knowledge management" has different meaning. We, computer science scientists, are concerned with the knowledge representation, data mining, and the knowledge structure that facilitates knowledge storage and retrieval with computers in mind. Thus, we will define the knowledge management as follows:

Knowledge Management (KM) comprises a wide range of methods/activities that extract information/knowledge from a body of unstructured raw data; organize the extracted information into structured form called knowledge; and design knowledge databases that are able to store and retrieve knowledge in an efficient way using computers.

In the above definition, we mentioned several terms such as raw data, information, and knowledge. What are the differences and the relationships among them? And the more fundamental question: how do we reason when faced with these entities? In the following sections, we will address these questions. First, we will outline the contributions of this document in the next section.

2. Contributions

In this chapter, we will introduce a computer reasoning method called "evidence theory" that is based on Bayes' theorem. We will describe relationships among raw data, knowledge, and information; we will implement a prototype of the evidence based reasoning software

component in the context of the bio-information system framework. The prototype is implemented with Java language and is applied to a medical case example: colorectal cancer. The evidence based reasoning theory proposed in this chapter will have significant impact on computer reasoning and artificial intelligence research.

With the increase of raw power in computer hardware, the search for better intelligent systems never ends. The research topics cover a wide range of areas. For example, some studies focus on the emotional aspect of an intelligent system (Fujita & et al., 2010), while others use statistical reasoning method in classifying news articles (Asy'arie, A. & Pribadi, A., 2009). Compare to existing literatures and reasoning methods, our presentation on the topic of computer reasoning (evidence based reasoning) is thorough. In addition, the reasoning method we proposed is generic in nature, thus it can be used in any domain. One key feature of our method is its simple calculation. Especially when number of evidence gets large, this simplicity becomes more important. Of course, you do not get this for free. You have to do the preparation by calculating degrees for evidences. But the saving you get is well worth the effort.

3. Background for data, information, knowledge, and wisdom

We can view data, information, knowledge, and wisdom as hierarchical in the context of knowledge management. With the data at the bottom and the wisdom at the top, we journey through concrete to abstract, and through no relationship to strong relationship as we move from left bottom to right top. This phenomenon is shown in Figure 1.

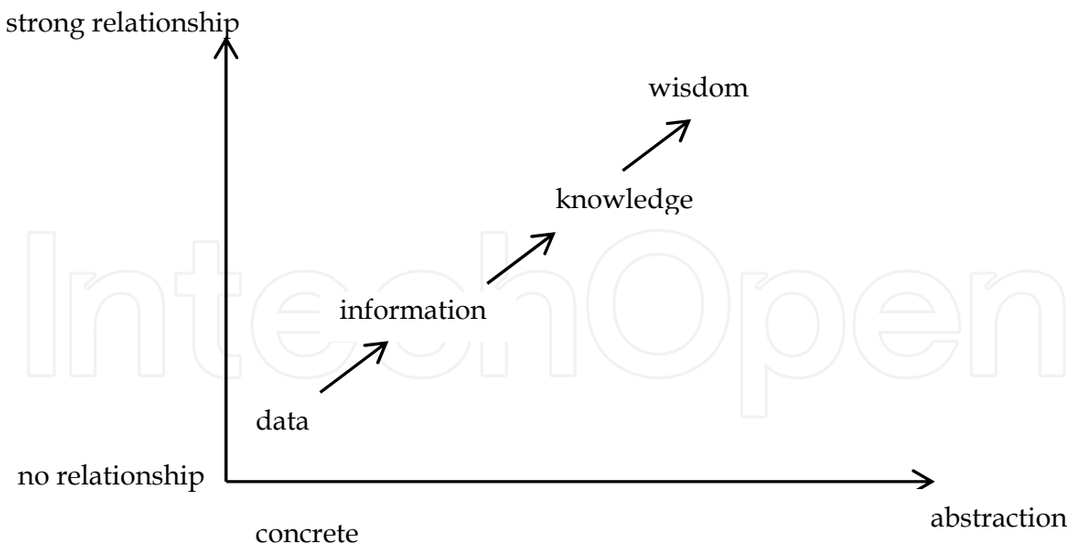


Fig. 1. Hierarchy of data to wisdom

Raw data are just meaningless points in data space. There are no references or relationships among these points. Raw data are like a phrase out of context. By themselves, they mean nothing. (referenced Bellinger, 2004)

As human, we often want to make sense out of raw data. When encountered a piece of data, we usually try to assign meaning to it, and try to find relationships for it. This is done by associating it with other things (or other data points). For example, consider the shapes in Figure 2.

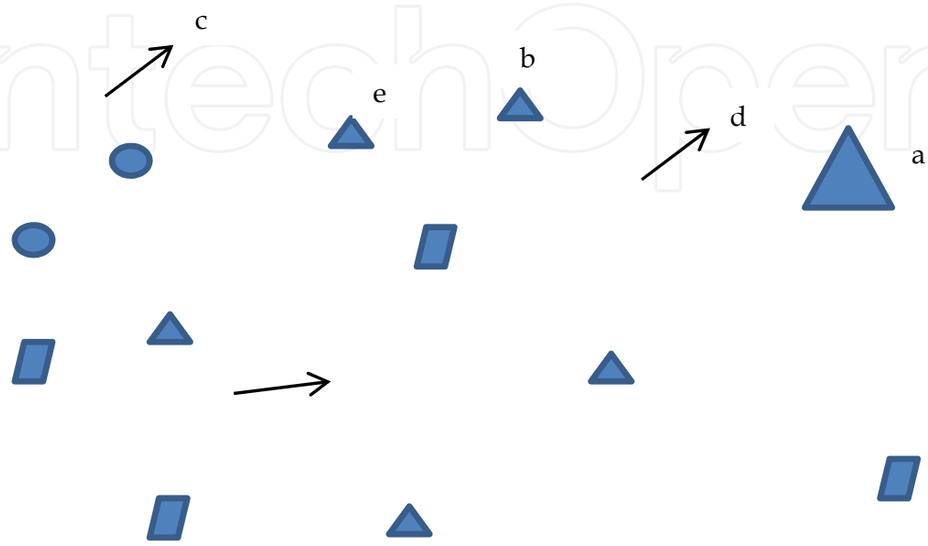


Fig. 2. Data points in data space

When seeing the items in Figure 2, we will automatically assign an "equal" relationship to item e and item b, assign the same relationship to item c and item d. We probably will assign a "similarity relationship" to item b and item a.

Take another example, if seeing the digits "3 4 5 …", most likely we will assign the meaning of "a sequence of positive integer numbers starting from 3 and extending to infinity." The point is that when there is no context, these raw data have no meaning. When we try to assign meanings to raw data, we are trying to create context for them. When raw data are put into context, some new things will happen.

The new things are information. There are some differences and relationships between raw data and information. First, information is not just a bunch of raw data piled together. Second, information is the interaction between the raw data and something we called "knowledge." Information depends on the understanding of the person perceiving the data. For example, the symbol "网 页"means nothing to an English speaker (to him, it is just raw data), but it conveys some information to a Chinese (to him, the same symbol is information and means a web page). The point is that whether some raw data represent meaningful information depends on the context. And the context is our prior experiences (often, we call these prior experiences "knowledge"). There is no guarantee that the information we extracted from raw data is correct. The correctness and usefulness of raw data depend on the knowledge of the person receiving the data. Another thing to point out is that the

experiences/knowledge will have influence on the interpretation of the data. The same piece of data may carry different meanings under different contexts (knowledge).

Knowledge is our prior experience. In other words, knowledge is the accumulated relationships and patterns that a person perceives among raw data. For example, if a layman sees the blood glucose test result of 230 mg/dL, he may have no clue as what this means. But for a trained doctor's eye, it means the person had the test is diabetic. The only difference here is the pattern. In the doctor's mind, from his prior training, a series of patterns such as:

Glucose level of 230 mg/dL -> diabetic

Diabetic -> risk of blindness

Diabetic -> risk of kidney failure

exist. On the other hand, there are no such patterns in the layman's mind. In essence, knowledge is the factoring of patterns (this includes summarization, abstraction, and crystallization of patterns). In the world of knowledge management in computer science, the knowledge is accumulated and crystallized patterns and relationships; and the information is the product of the interaction between data and knowledge. In other words, when connecting the dots, you are producing information.

Wisdom is the highest form of deep patterns. Usually, we only attribute wisdom to intelligent beings. Bellinger (2004) has the following description about wisdom:

"Wisdom arises when one understands the foundational principles responsible for the patterns representing knowledge being what they are. And wisdom, even more so than knowledge, tends to create its own context. I have a preference for referring to these foundational principles as eternal truths, yet I find people have a tendency to be somewhat uncomfortable with this labeling. These foundational principles are universal and completely context independent. Of course, this last statement is sort of a redundant word game, for if the principle was context dependent, then it couldn't be universally true now could it?"

In this documentation, we will focus on data, information, and knowledge. We will leave the topic of wisdom to philosophers. Particularly, we will deal with computer reasoning and knowledge management using knowledge databases. Before presenting our methods for the knowledge representation, reasoning, and knowledge management, we need to answer the philosophical question: is there any difference between human reasoning and computer reasoning? Our answer is "Yes."

Computer reasoning and human reasoning are different. One of the biggest differences has something to do with creative ideas. Often, we see someone with so called "killer ideas." "Killer ideas" refer to those ideas that are revolutionary, creative, and not conform to the norms of the contemporary generation. For example, Sir Isaac Newton's law of gravity, Albert Einstein's theory of relativity, and the idea of ten dimensional UNIVERSE are all examples of killer ideas. How exactly these "killer ideas" are produced is still open for debate. However, we do know computers are incapable of producing these ideas (at least for the time being); because, we haven't seen any computer that can produce any meaningful killer ideas yet. Thus, we conclude that computers reasoning and human reasoning are

different. With current technology, we can delegate computers to reason with low level entities (in the low-left of Figure 1) in the knowledge management hierarchy. This is because at lower levels, reasoning is more objective and concrete. The theoretical foundations of the reasoning at lower levels can be captured by the Bayes' theorem.

4. Theoretical foundation of computer reasoning

The insight that we get from the above discussion on raw data, information, knowledge, and wisdom tells us that reasoning at lower levels is easier than reasoning at the highest level. Since at the lower levels, we only need to deal with knowledge finding (data mining) and the application of the appropriate knowledge to some evidences. At the highest level (the killer idea level), we even do not know the mechanism that produces creative ideas; therefore, it will be much harder to reason at this level. As mentioned in the previous section, the theoretical foundation of computer is Bayes' theorem. Let's investigate what is the Bayes' theorem? Bayes' theorem can be expressed as Formula 1:

$$P(A|X) = \frac{\mathrm{P(X|A)} * P(A)}{\mathrm{P(X|A)} * P(A) + P(X|{\sim}A) * P({\sim}A)} \quad \text{(Formula 1)}$$

The notation P(A | X) means the probability (or the chance) that the event A will happen given the evidence (or the observation) of X. In probability theory, this is called conditional probability. Depending on the quality of evidence X, the probability of event A happening may be heavily affected by the presence of the evidence X.

The symbol "~" means complement, that is, the opposite of what follows it. For example, if P(A) means the probability of event A will happen, then P(~A) means the probability of event A will not happen. One thing to point out is that there are three pieces in Formula 1: the reasoning about the occurrence of an event A (the left side of the equation), the evidence (X), and the causality relationship between the evidence X and the event A (embodied by P(X | A) and P(X | ~A)).

In a nutshell, Formula 1 says that if we see a piece of evidence X, we can reason about the chance of event A's occurrence given that the evidence X and the event A has a causality relationship. This is exactly the behavior that a rational person will display given a piece of evidence related to the event. Formula 1 can be extended to include two, three, …, and many pieces of evidence. All we need to do is to apply the formula multiple times. For example, if both X and Y contribute to the occurrence of event A, we can calculate the final probability of event A by applying Formula 1 to get the probability of A given evidence X. Then, we use the result to apply Formula 1 again. Only this time, we should use the result in the first iteration to substitute the prior probability P(A), and P(~A). Actually, we can repeatedly apply Formula 1 to reason any number of evidences.

To get a better handle on how the Bayes' theorem works, let's work through a concrete example. Suppose that we have the following problem statement:

Example 1: "Lung cancer is the leading cause of cancer death in the United States." (Williams, 2003, p. 463) Suppose that about 0.2% of the population living in US with age above 20 has lung cancer. When doing an annual check, suppose that 85% of the people with lung cancer will show positive for the chest x-ray test. On the other hand, chest x-ray will have false alarms: 6% of the people without lung cancer will also show positive for the chest

x-ray test. If a person went through the annual check and had a positive chest x-ray, what is the probability that he/she has the lung cancer? (For concreteness, you may assume that there are 10,000 people participated the annual check)

Answer: Most people will give the wrong answer of "the person will have 85% probability of having the lung cancer."

To get the answer right, we must first understand several important facts in statistics. The first thing is that

$$P(A \mid X) \neq P(X \mid A)$$

The reason that most people will get the incorrect answer of 85% is the confusion caused by the above inequality relationship.

The correct answer for Example 1 is 2.8%. The following is the analysis and steps showing how we get the correct answer:

1. We start out by the basic probability definition:

   P(cancer | positive x-ray) = number of people who have both cancer and positive x-ray in the annual check / total number of people with positive x-ray in the annual check (Formula 2)

   According to the meaning of conditional probability, the left side of Formula 2 is the answer we are looking for.

   Note: the key of the above equation is to use the number of people who have both cancer and positive x-ray as the numerator. If using people who have cancer as the numerator, the result will be wrong since there are people who have cancer but have negative x-ray test results.

2. We use concrete number. Without losing generality, we assume there are 10,000 people of age 20 and over participated in the annual check. Thus, we have the following data:

   The number of people who have lung cancer in the annual check group is 10,000*0.2% = 20.

   The number of people who are healthy in the annual check group is 10,000*99.8% = 9980.

   The number of people who have lung cancer and have positive x-ray is 20*85% = 17.

   The number of people who have lung cancer and have negative x-ray is 20*15% = 3.

   The number of people who have no lung cancer and have positive x-ray is 9980*6% = 599.

3. We use the data in step 2 and plug into the Formula 2 in step 1. We will get following answer:

   P(cancer | positive x-ray) = 17 / (17+599) = 17 / 616 = 0.028

Most people regard Bayes' theorem as statistical formula and overlook its reasoning logic. We want to point out that it is also a reasoning method that captures the essence of reasoning logic that reasons at the lower-levels. Thus, it is the theoretical foundation that underpins the computer reasoning.

5. Bayes' reasoning

Example 1 in previous section can also be solved by Bayes' theorem. One thing to remember in understanding Bayes' theorem is the following statistical formula:

$$P(A \& B) = P(A \mid B) * P(B) \qquad \text{(Formula 3)}$$

Or equivalently,

$$P(A \& B) = P(B \mid A) * P(A)$$

In the following, we will explain how Bayes' reasoning works and the meaning of its subparts.

5.1 Bayes' reasoning explained

Bayes' theorem can be viewed as the bridge that connects the reasoning to physical evidences: on the left of Formula 1 is the inference/reasoning, and on the right of Formula 1 is the physical evidence that supports the reasoning on the left. When estimating the prior probabilities (prior probability includes: the baseline probability P(A), the two conditional probabilities: P(positive x-ray | cancer) and P(positive x-ray | healthy)) on the right, we are constructing a reasoning model; when applying the Bayes' theorem, we are extracting information using the constructed model. The process of applying the theorem is the process of combining the raw data, the knowledge (context) to yield information. We can use the Bayes' theorem to solve Example 1 as follows:

1. Start with what we want to achieve: P(cancer | positive x-ray)
2. Rewrite it as following with the help of Formula 3:

$$P(\text{cancer} \mid \text{positive x-ray}) = P(\text{cancer} \,\&\, \text{positive x-ray}) \,/\, P(\text{positive x-ray})$$

3. P(positive x-ray) can be expanded to P(positive x-ray & cancer) + P(positive x-ray & ~cancer).

Note: this expansion captures the causality of the reasoning model. It says that the total number of people with positive x-ray in the annual check group is coming from two groups: the people with cancer and show the positive x-ray and the people with no cancer and show the positive x-ray.

4. Plug in the result from step 3 to the equation in step 2, we get

$$P(cancer|positive\ x-ray) = \frac{\mathrm{P}(\text{cancer} \,\&\, positive\ x-ray)}{\mathrm{P}(\text{positive x-ray} \,\&\, cancer) + P(positive\ x-ray \& \sim cancer)}$$

5. The above equation can be rewritten as following with the help of Formula 3:

$$P(cancer|positive\ x-ray) = \frac{\mathrm{P}(positive\ x-ray \mid cancer) * P(cancer)}{\mathrm{P}(positive\ x-ray \mid cancer) * P(cancer) + P(positive\ x-ray|\sim cancer) * P(\sim cancer)}$$

This is exactly the same formula as in the Bayes' theorem (Formula 1). If we use the following data implied by the problem statement in Example 1:

P(cancer) = 0.2% (20 out of 10,000 have cancer) (1)

P(~cancer) = 99.8% (9980 out of 10,000 have no cancer) (2)

P(positive x-ray | cancer) = 85%

(85% of people with lung cancer have positive x-ray) (3)

P(positive x-ray | ~cancer) = 6%

(6% of people without lung cancer have positive x-ray) (4)

And plug in the above data into the above expression, we will get:

P(cancer | positive x-ray) = 85%* 0.2% / (85%*0.2%+6%*99.8%)

= 0.0017 / (0.0017+0.06)

= 0.0017 / 0.0617

= 0.028

This is exactly the same answer we got in the previous section.

Bayes' reasoning needs three pieces of information (all appear on the right of the equation at the beginning of step 5): the percentage of people with lung cancer, the percentage of people without lung cancer who have false alarms, and the percentage of people with lung cancer who show positive on the test. The first piece of information which is part of the priors is the baseline knowledge. The second and third pieces of information which also belong to the priors are the measurement of the quality of evidence. Bayes' reasoning is to use the evidence to change the belief/knowledge (shifting the baseline upwards with positive evidence or downwards with negative evidence). We will use more examples to show how this change of belief (the machine reasoning) happens. The left-side probability is the posterior probability. It is the revised view of the world in the light of evidence which is on the right-side of the equation.

To see how the first piece of information affects the Bayes' result, let's assume that the batch of people doing the annual check is high risk smokers. According to Williams (Williams, 2003, p. 464), smoker's chance of getting lung cancer is 13 times higher than non-smokers. Now, let's ask the same question: what is the probability of the person has lung cancer if he/she has the positive x-ray test given that the cancer rate in this group is 2.6% (2.6% is getting from 0.2* 13)? Sure enough, the final answer should be different. Actually, the new answer is 27.4%. The following is the analysis and steps showing how we get the correct answer:

1. We use the Bayes' theorem:

$$P(cancer|positive\ x-ray) = \frac{\mathrm{P}(positive\ x-ray \mid cancer) * P(cancer)}{\mathrm{P}(positive\ x-ray \mid cancer) * P(cancer) + P(positive\ x-ray|\sim cancer) * P(\sim cancer)}$$

2. And plug in the following data:

$$P(\text{cancer}) = 2.6\% \quad (260 \text{ out of } 10{,}000 \text{ have cancer}) \tag{5}$$

$$P(\sim\text{cancer}) = 97.4\% \quad (9740 \text{ out of } 10{,}000 \text{ have no cancer}) \tag{6}$$

$$P(\text{positive x-ray} \mid \text{cancer}) = 85\%$$

$$(85\% \text{ of people with lung cancer have positive x-ray}) \tag{7}$$

$$P(\text{positive x-ray} \mid \sim\text{cancer}) = 6\%$$

$$(6\% \text{ of people without lung cancer have positive x-ray}) \tag{8}$$

3. And plug in the above data into the above Bayes' theorem, we will get:

$$\begin{aligned} P(\text{cancer} \mid \text{positive x-ray}) &= 85\%* 2.6\% / (85\%*2.6\%+6\%*97.4\%) \\ &= 0.0221 / (0.0221+0.0584) \\ &= 0.0221 / 0.0805 \\ &= 0.274 \end{aligned}$$

As you can see, comparing to the non-risky population (the probability of having cancer 0.028), the probability value of 0.274 of a person in the risky group is much higher. This makes sense since the prior probability of getting lung cancer is higher in this high risk group. In this new example, the quality of the x-ray equipment does not change. The only thing changed is the prior cancer rate, from 0.2% to 2.6%. At first look to the new problem, most people will give the same wrong answer of 85%. But Bayes' reasoning gives us more objective and correct answer. Here is an example that computer reasoning can be better than a human!

Bayes' reasoning can be used in situations that have multiple evidences. Let's use Example 2, which is the extension of Example 1, to illustrate how this is done.

Example 2: "Lung cancer is the leading cause of cancer death in the United States." (Williams, 2003, p. 463) Suppose that about 0.2% of the population living in US with age above 20 has lung cancer. When doing an annual check, assume that 85% of the people with lung cancer will show positive for the chest x-ray test. On the other hand, chest x-ray will have false alarms: 6% of the people without lung cancer will also show positive for the chest x-ray test. Suppose that a hospital will do two lung cancer screen tests for each annual check patient (assume the two tests are independent). The second test called CT scan is done to improve the accuracy of diagnosis. Suppose that the CT scan has the following characteristics: it returns positive for 85% of the people with lung cancer; it has a lower false rate than the x-ray test and will return false positive for one out of one thousand people without lung cancer. If a person went through the annual check and had positives on both the chest x-ray and the CT scan, what is the probability that he/she has the lung cancer?

Answer: We can solve this problem by using the Bayes' theorem twice. We already know that the probability of a person has cancer given that he has positive x-ray is 2.8%; the probability of a person has no cancer given that he has positive x-ray is 97.2%. We can use this result and continue to solve the problem as follows:

1. We use the Bayes' theorem:

$$P(cancer \mid positive\ x-ray\ \&positive\ CT\ scan) = \frac{\mathrm{P}(positive\ CT\ scan \mid cancer) * P(cancer\ new\ proir)}{\mathrm{P}(positive\ CT\ scan \mid cancer) * P(cancer\ new\ prior) + P(positive\ CT\ scan|{\sim}cancer) * P({\sim}cancer\ new\ prior)}$$

2. And plug in the following data:

P(cancer new prior) = 2.8% (the poster probability of) (9)

P(~cancer new prior) = 97.2% (the complement of equation (9)) (10)

P(positive CT scan | cancer) = 85%
(85% of people with lung cancer have positive CT scan) (11)

P(positive CT scan | ~cancer) = 0.1%
(0.1% of people without lung cancer have positive CT scan) (12)

3. And plug in the above data into the above Bayes' theorem, we will get:

P(cancer | positive x-ray & positive CT scan) =

85%* 2.8% / (85%*2.8%+0.1%*97.2%)

= 0.0238 / (0.0238+0.00097)
= 0.0238 / 0.02477
= 0.96

As you can see, the person's probability of having lung cancer is very high in this instance. In this example, each application of the Bayes' theorem can be viewed as a mapping from one statistical sample space to another statistical sample space and there are two such mappings as shown in Figure 3.

In Figure 3, P(xp&c) means the probability of a person who has lung cancer and is x-ray positive; P(xp&CTp&c) means the probability of a person who has lung cancer and is both x-ray positive and CT scan positive. Similarly, P(xp&CTp&h) means the probability of a person who is healthy and is both x-ray positive and CT scan positive. To help our understanding of what's going on, we list some calculated data below (assume total of 10,000 people):

Prior probability P(cancer) = 0.2% number of cancer people=20
Prior probability P(healthy) = 99.8% number of healthy people=9980
conditional probability P(positive x-ray | cancer) = 85%
number of people having cancer and positive=P(p|c)*#cancer=17
conditional probability P(positive x-ray | healthy) = 6%
number of people who are healthy and positive=P(p|h)*#healthy=599

P(cancer | positive x-ray) = #people having cancer&positive / total#people having positive
= 17 / (17 + 599) = 0.028

posterior probability (our answer)

As shown in Figure 3, the application of Bayes' theorem is the mapping from one space to another space. In the initial world, the probability of a person in the sample is healthy is 99.8% while the probability of having the lung cancer is 0.2%. The first application of the Bayes' theorem has two distortions: one distorts the probability of having cancer, P(cancer), to the probability of both having cancer and being positive for x-ray test, P(positive x-ray & cancer), (the distorting leverage/filter is the conditional probability P(positive x-ray | cancer)), the other distorts the probability of being healthy, P(healthy), to the probability of being positive for x-ray and being healthy, P(positive x-ray & healthy), (the distorting leverage/filter is the conditional probability P(positive x-ray | healthy)). In the new alternate universe, though the number of people who have cancer to be included is almost the same as in the initial world (from 20 in the initial world to 17 in the first mapped world), the number of people who are healthy to be included is greatly reduced (from 9980 in the initial world to 599 in the first mapped world). Thus, when we try to answer the question of "the probability of a person having lung cancer given that he has a positive x-ray" by dividing the number of people with cancer by the total number of people with positive x-ray, we will get a much higher probability. In other words, the mapping altered our assessment. The mapping reflects the effect of the evidence "positive x-ray" in shifting our judgment of deciding whether a person has lung cancer.

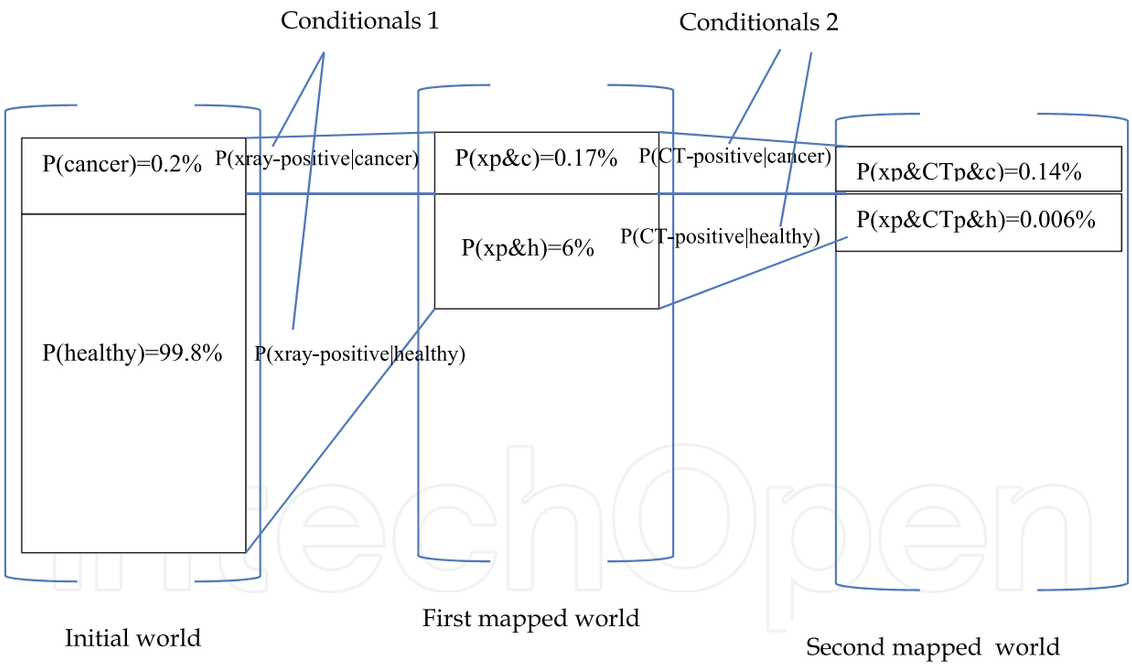


Fig. 3. We apply Bayes' theorem twice for two tests; each application of Bayes' theorem can be viewed as a mapping

One thing to point out, the x-ray test will not affect the actual probability of a person has cancer (otherwise no one will take the test). However, the test will affect our beliefs. A positive x-ray is a membership test. If the test is positive, it will eliminate many more people without lung cancer than people with the cancer. The number of people without cancer is reduced by a factor of more than 16, from 9980 to 599, while the number of people with

cancer is reduced only from 20 to 17. Thus, the proportion of 17 within 616 (the total number of people with positive x-ray) is much larger than the proportion of 20 within 10,000.

5.2 Conditional probabilities play the role as shifters

From Example 2, you may have already seen the role played by the two conditional probabilities: P(positive x-ray | cancer) and P(positive x-ray | healthy). They are the shifters: P(positive x-ray | cancer) shifts our view positively and P(positive x-ray | healthy) shifts our view negatively. In other words, large value of P(positive x-ray | cancer) will increase our confidence in predicting a person has cancer given that he has a positive test. On the other hand, small value of P(positive x-ray | healthy) will increase our confidence in predicting a person has cancer given that he has a positive test. The quality measurement of a test in altering our view to the world is the inter-play of these two conditional probabilities. They map the number of cancerous people and the number of healthy people in one world into another world. Their ratio can be used as a measurement of effectiveness for a test to be evidence.

We will show later that for a test to be effective, its positive conditional probability cannot have the same value as its negative conditional probability. Otherwise, the test will shift our view to the same amount and the net effect is nil.

The second application of the Bayes' theorem alters the ratio of number of healthy people to the number of cancer people in the universe even further. In the second mapped world, the number of people who have cancer to be included is 14, and the number of people who are healthy to be included is 0.6. In the second new world, seeing both positive evidences (a positive x-ray and a positive CT scan) is convincing evidence that the person has lung cancer (96% probability).

Bayes' theorem is important in understanding the basic statistical reasoning mechanism. In its original form, it is not easy to use, especially in the face of multiple evidences. In the next section, we will introduce a computer reasoning theory: evidence theory that is based on the Bayes' theory.

6. The evidence theory of computer reasoning

In this section, we are going to present a computer reasoning method called evidence theory that is more convenient and easier to use than the Bayes' theorem. To help our presentation, we will define some terms and use some mathematical formulas along the way.

If we take an abstract view, the computer reasoning can be summarized as: capture the causality relationships from raw data, build a knowledge database using these relationships, and make a judgment (or inference) on pieces of evidence based on existing knowledge database. The essence of the summary is shown in Figure 4.

The computer reasoning mechanism shown in Figure 4 can be explained as having two stages: the knowledge/pattern building and the application of knowledge to the new evidence. In the first stage (indicated by arrows from the Raw data to the Knowledge database), knowledge is produced either by data mining from raw data or by direct human insertion; in the second stage (indicated by arrows from the Knowledge database to the

Solution), the reasoning occurs by applying the knowledge from the Knowledge database to the evidence.

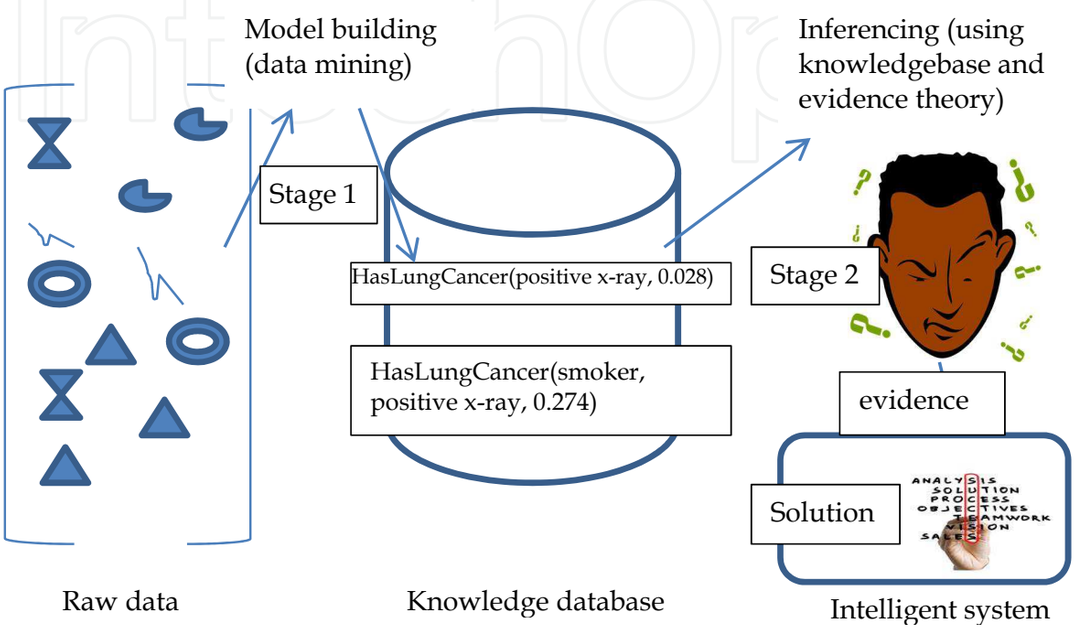


Fig. 4. Abstract view on computer reasoning

One of the important components in Figure 4 is evidence. In computer reasoning, evidence is the main factor that influences a computer's judgment. One of the important characteristics of evidence is its quality. In this abstract view, reasoning is persuaded by the presence of evidence. For example, without any evidence, our view of the initial world about the probability of a person with lung cancer is 0.2%, with the presence of first piece of evidence, the positive result of x-ray test, our view of the modified world about the probability of a person with lung cancer is 2.8%, with the presence of two pieces of evidence, the positive result of x-ray and the positive result of CT scan, our view of the new world about the probability of a person with lung cancer is changed again to 96%.

6.1 The properties and the definition of evidence (or a test)

The main role of a piece of evidence is its influence on a rational mind. To see how this influence is realized, we need to investigate the properties about evidence. In this section, we will give definition of evidence; and will describe properties of evidence. These definition and properties are given in the following highlight box.

Evidence Theory and Evidence Properties

The Main Interest: Suppose that A represents an event of interest; E represents the a piece of envidence. The main interest of the evidence theory is to calculate the probability:

$$P(A \mid E) \qquad \text{(Formula 4)}$$

Definition of Evidence: we define evidence (or a test) E as a piece of information that has the ability to change the value of probability defined in Formula 4. The underlining reason for this ability is the causality relationship existed between the event A and the evidence E.

Assumption about Event A: in the absence of any evidence, we will assume that the probability of event A occurring is the same as the probability of its not occurring. That is, P(A) = 50%.

Properties of Evidence: evidence has following three properties:

Property 1: if evidence E increases the probability of event A, then the evidence E is positive evidence relative to event A.
Property 2: if evidence E decreases the probability of event A, then the evidence E is negative evidence relative to event A.
Property 3: the quality of evidence E is measured in terms of evidence strength (which will be defined in the next section).

6.2 The quality of evidence (evidence strength)

As mentioned before, one important function of a piece of evidence is its influence on a rational mind. Thus, the quality measurement of a piece of evidence should also be based on its ability to influence. For example, if evidence A convinced us an event (or goal achievability) will happen with 80 percent certainty while evidence B convinced us the same event will happen with 90 percent certainty, then we would say evidence B is better. We can quantify the quality of evidence by introducing the concept of evidence strength. With this measurement criterion in mind, try to answer the following question:

Question 1: With regard to the two tests mentioned in Example 2: the x-ray test, and the CT scan test, which one is better in swaying us to believe that the person in question has lung cancer?

Here is the repeat of some statistics for the two evidences (a medical test can be regarded as evidence from Bayes' theorem's point of view):

- X-ray test: 85% of the people with lung cancer will show positive; 6% of the people without lung cancer will also show positive.
- CT scan test: 85% of the people with lung cancer will show positive; 0.1% of the people without lung cancer will also show positive.

Before answering the above question, let's define some terms. In the following, "Posi|Cause" means that the existence of "Cause" causes the evidence "Posi" to appear; "Posi|~Cause" means that the absence of "Cause" causes the evidence "Posi" to appear. Now, we will define the strength of evidence as follows:

Definition of evidence strength: we define strength of evidence (or a test) as the probability that the evidence gives true positive divided by the probability that the evidence gives a false positive. In other words, it can be represented as the following formula:

$$\text{strength(evidence)} = P(\text{Posi}|\text{Cause}) / P(\text{Posi}|\sim\text{Cause}) \qquad \text{(Formula 5)}$$

One thing to point out is that the summation of the probability of P(Posi|Cause) and the probability of P(Posi|~Cause) is not necessarily 1. Once defined evidence strength, we can divide evidence two categories: positive evidence and negative evidence. When the value of strength is greater than 1, the evidence will shift our belief in the positive way, thus we name it positive evidence; on the other hand, when the value of strength is smaller than 1, the evidence has the effect of shift our belief in the negative way, thus we name it negative evidence.

The probability P(Posi | Cause) on the right side of Formula 5 captures the causality relationship in the real world. It means the probability of something causes the evidence (test) to be positive. In our Example 1, it will take the form: P(positive x-ray | cancer), and it means that the probability of lung cancer causes the x-ray to be positive; and P(positive x-ray | ~cancer) means the probability of a false alarm.

Now, let's give some observations about evidence. First, as mentioned before, to be effective evidence, the value of a test's positive conditional probability cannot have the same value as its negative conditional probability. Thus, in terms of strength, we have the following observation:

Observation 1: when the evidence strength is 1, it is not good evidence. Using the above definition, the effectiveness of a test (or a piece of evidence) is measured in terms of its strength. If the value of strength is 1, then the test is useless as a piece of evidence (it is neutral). When the value of strength is greater than 1, it is positive evidence (seeing the evidence will shift our view regarding the trueness of the event "Cause" to the positive side); when the value of strength is smaller than 1 and greater than 0, it is negative evidence (diminishes our view about the trueness of the "Cause").

For example, if we are asked whether flipping a fair coin is a good test for predicting a person has lung cancer (assume that a head means the person has cancer and a tail means the person has no cancer)? We can proceed like the following:

1. First, we calculate the strength of flipping a coin as a test and it will be:

$$\text{strength(flipping a coin)} = P(\text{head} \mid \text{cancer}) / P(\text{head} \mid \sim\text{cancer}) = 0.5 / 0.5 = 1$$

Note: the reason that P(head | cancer) = 0.5 is the fact that the information of a patient has cancer has nothing to do with the outcome of flipping a coin. The chance of getting a head is still governed by its old chance of 50%. We will have the same argument for the probability P(head | ~ cancer).

2. Based on our evidence theory, we know it shifts our belief to the same distance for positive and negative direction. Thus, we conclude that it's not a good test.

With regard to the cause of strong evidence, we have the following observation:

Observation 2: strong evince is not caused by a very high probability of cause leads to the positive test, rather it is caused by a very low probability of not-cause could have led to the positive test.

For example, if it is raining, the grass in my front yard (there is no roof) is likely to be wet. But seeing the grass wet does not necessarily mean that it is raining (maybe it is caused by the sprinkler). In other words, when seeing the evidence of wet grass, we cannot reason that it is raining with certainty. This is a case of high probability of cause-effect but week evidence.

On the hand, if we are watching an area there is no sprinkler. Then, seeing the wet grass would always mean that it is raining, even though we assume that there is a weak causation link such as the rain will cause the grass wet only 60% of times. This is a case of low probability of cause-effect but strong evidence.

Now, let's answer the Question 1. We will use the evidence strength value to help us make the conclusion. For x-ray test, we have the following:

$$\text{strength(x-ray test)} = P(\text{positive x-ray} \mid \text{cancer}) / P(\text{positive x-ray} \mid \sim\text{cancer})$$
$$= 0.85 / 0.06 = 14.17$$

For CT scan, we have:

$$\text{strength(CT scan test)} = P(\text{positive CT scan} \mid \text{cancer}) / P(\text{positive CT scan} \mid \sim\text{cancer})$$
$$= 0.85 / 0.001 = 850$$

Since the value 850 is greater than 14.17, we conclude that CT scan test is a better evidence in convincing us that the patient in question has lung cancer.

6.3 The relationship between the evidence strength and its influence power

The discussion above gives us some insights about evidence. In this section, we will investigate the relationship between the evidence strength and its power to influence the outcome of an event. Specifically, we want to see how the existence of a piece of evidence will shift our belief (its direction and its amount (may be rough estimation)). Based on the intuition we have about the evidence, we make the following claim.

Claim 1: the influence power of a given piece of evidence is proportional to the value of evidence strength. For positive evidence, the larger evidence strength value, the stronger the influence power; for negative evidence, the smaller evidence strength value, the stronger the influence power.

We will use the following example to give some insight about our Claim 1:

Example 3: Using the data in Example 2, calculate the strength for x-ray test and the strength for the CT scan test. Then, calculate the distance that each test moves our belief (including the direction) in terms of percentage change. We repeat the main points and data in the following:

1. About 0.2% of the population living in US with age above 20 has lung cancer.

2. When doing an annual check, assume that 85% of the people with lung cancer will show positive for the chest x-ray test. About 6% of the people without lung cancer will also show positive for the chest x-ray test.
3. The second test called CT scan is done independently. It returns positive for 85% of the people with lung cancer; its false rate is 0.1%.

Answer: For first part of the question, we can use the result in the previous section. Here is the repeat: For x-ray test, we have the following:

$$\text{strength(x-ray test)} = P(\text{positive x-ray} \mid \text{cancer}) / P(\text{positive x-ray} \mid \sim\text{cancer})$$
$$= 0.85 / 0.06 = 14.17$$

For CT scan, we have:

$$\text{strength(CT scan test)} = P(\text{positive CT scan} \mid \text{cancer}) / P(\text{positive CT scan} \mid \sim\text{cancer})$$
$$= 0.85 / 0.001 = 850$$

For the second part of the problem (the distance each test sways our beliefs), we will proceed as follows:

We started in the initial world with following probabilities:

$$P(\text{cancer}) = 0.2\%,\ P(\text{healthy}) = 99.8\%$$

For a person in this initial world, the probability of having lung cancer is 0.2% (pretty low). If we use the x-ray as a membership test, then the probability become following (already calculated in previous sections):

$$P(\text{cancer} \mid \text{positive x-ray}) = 2.8\%,\ P(\text{healthy} \mid \text{positive x-ray}) = 97.2\%$$

The x-ray test shifted our view from P(cancer) = 0.2% to P(cancer | positive x-ray) = 2.8%. It is a positive evidence. The percentage increase is 2.6%.

Now, let's see how much the CT scan test will shift our view. Starting from the initial world, if we use the CT scan as a membership test, then the probability can be calculated as following:

We use the Bayes' theorem:

$$P(cancer|positive\ CT\ scan) = \frac{P(positive\ CT\ scan \mid cancer) * P(cancer)}{P(positive\ CT\ scan \mid cancer) * P(cancer) + P(positive\ CT\ scan|\sim cancer) * P(\sim cancer)}$$

And plug in the following data:

$$P(\text{cancer}) = 0.2\% \quad (20 \text{ out of } 10{,}000 \text{ have cancer})$$

$$P(\sim\text{cancer}) = 99.8\% \quad (9980 \text{ out of } 10{,}000 \text{ have no cancer})$$

$$P(\text{positive CT scan} \mid \text{cancer}) = 85\%$$

(85% of people with lung cancer have positive CT scan)

P(positive CT scan | ~cancer) = 0.1%

(0.1% of people without lung cancer have positive CT scan)

And plug in the above data into the above Bayes' theorem, we will get:

P(cancer | positive CT scan) = 85%* 0.2% / (85%*0.2%+0.1%*99.8%)
= 0.0017 / (0.0017+0.001)
= 0.0017 / 0.0027
= 0.63

This result tells use that the CT scan test will shift our belief in positive direction, The percentage increase is 62.8%. These results support our claim 1.

Note that the x-ray test and CT scan test have the same positive cause-effect probability rate but different false alarm rate. In x-ray test, the false alarm probability P(positive x-ray | ~cancer) is 6%, while in CT scan test, the false alarm probability P(positive CT scan | ~cancer) is only 0.1%. Here is an example that the **low false alarm probability is the dominate factor** in deciding the strength of evidence.

6.4 The logarithmical representation of evidence degrees

In the previous section, we used the ratio of two conditional probabilities as the strength measurement. Under our abstract view of reasoning model in Figure 4, evidences are used to distort the world space. As indicated in that figure, reasoning is the process of make a judgment using the knowledge (embedded in the conditionals) based on the evidence (the right side of "|" on the left side of Formula 1) presented. One thing to point out is that our abstract reasoning model can be applied to multiple evidences.

To capture the essence of the low-level reasoning in situations with multiple evidences, we can use a tool in mathematics called ratio and the concept in statistics called odds. Also, the use of these tools will make reasoning in situations that have multiple evidences easier. Odds can capture the same information as probability. In statistics, odds are defined as the ratio of the probability of an event's occurring to the probability of its not occurring. The reasoning of solving the problem in Example 2 using the odds concept will be like this: in the initial world, the lung cancer rate is 0.2%. Thus, 2 out of 1000 people have lung cancer, and 998 people out of 1000 people do not have lung cancer. Using odds, we define the event of interest as a person has lung cancer vs. a person has no cancer. So the 0.2% cancer rate can be expressed as the following odds:

2:998

And the evidence strengths of the two tests x-ray, and CT scan can be expressed in odds notation as:

14.17:1 (get from 0.85/0.06)

850:1 (get from 0.85/0.001)

To get the answer for low level reasoning, we calculate the odds for a person with cancer who score positive on the two tests, versus a person without cancer who score positive on the two tests. Using the basic principles in algebra, the above odds can be calculated as following

2*14.17*850:998*1*1 =

24089:998

Once get the final odds, we can get probability of a person having lung cancer given that he score both tests positive as following:

P(cancer | positive x-ray & positive CT scan) = 24089 / (24089+998)

= 24089 / 25087

= 96%

This is the same answer as we get using Bayes' theorem in section 5.

As you can see, using the ratio and the odds tool is simpler than using the Bayes' theorem directly. We can simplify our calculation even further by using another tool called logarithm in mathematics. Before we can take the advantage of logarithm, we need to give a new definition on evidence called evidence degree.

Definition of evidence degree: we define evidence degree of a test as the as the following formula:

$$\text{degree(test)} = 10 \log_{10} \text{strength (test)} \quad \text{(Formula 6)}$$

To get the strength from the degree, we use the following formula:

$$\text{strength(test)} = 10^{\text{degree(test)} / 10} \quad \text{(Formula 7)}$$

Once represented in logarithmic format (degree of evidence), the aggregated effect of evidence toward a goal can be obtained by simple adding instead of multiplying.

6.5 The evidence based reasoning

As mentioned before, at low-level reasoning, the logic employed by a human is the same as the Bayes' theorem. In this section, we will show how to reason using the evidence expressed in the form of degree. As the topic suggested, the focus of our reasoning method is on evidence. The reasoning method addresses the question of the following type:

Question Type: Given a set of evidences and prior probability of an event A, we want to reason about the posterior probability of A (here, the event of interest can be anything, such as the survival chance of a disease, the goal in a planning problem, etc.). In other words, we want to figure out the left side of the following equation:

P(A | seen evidences x, y, z, …) = ?

The assumption of this method is that each piece of evidence is independent. Because of the strength of Bayes' theorem, this assumption works even for evidences that are not independent. Studies show that systems based on Bayes' theorem with the same assumption such as Hidden Naïve Bayes (Jin & et al., 2007) are robust because of the model constructing can accommodate minor factors easily. The reason for this robustness stems from the fact that the model itself has already captured the main causality. Any other accuracy consideration does not improve too much. In a sense, it only adds the complexity.

Our reasoning method can be represented as the following algorithm:

EvidenceBasedReasoning Algorithm: Inputs: raw data, input question of probability of an event of interest; Output: posterior probability information (answer to the input question)

Step 1: constructing models (or knowledge) from raw data.
Step 2: calculate the quality of evidence related to the input question in terms of evidence degree with the help of Formulas 6 and 7.
Step 3: calculate the overall evidence degree.
Step 4: interpret the information by converting the overall evidence degree back to the probability (using Formulas 6 and 7 again).

We have the following comments about the degree of evidence:

1. The critical point for the degree of evidence is 0. 0 means the evidence is neutral; the probability of positive conditional is equal to the probability of negative conditional. It does not add anything in shifting our view to the world.
2. If the evidence's degree is greater than 0, then it will shift our view toward believing event A is true; if the evidence's degree is less than 0, then it will shift our view toward believing event A is not true;
3. Degree is measured in terms of order of degree. If evidence A's degree is 10 and evidence B's degree is 20, then evidence B is ten order of magnitude (100 times) stronger than evidence A in persuasion power.

Now, let's use an example to illustrate how our EvidenceBasedReasoning works.

Example 4: Solve the problem in Example 2 again using the EvidenceBasedReasoning algorithm. We repeat the main points and assumptions in the following:

1. About 0.2% of the population living in US with age above 20 has lung cancer.
2. When doing an annual check, assume that 85% of the people with lung cancer will show positive for the chest x-ray test. About 6% of the people without lung cancer will also show positive for the chest x-ray test.
3. The second test called CT scan is done independently. It returns positive for 85% of the people with lung cancer; its false rate is 0.1%.
4. If a person went through the annual check and had positives on both the chest x-ray and the CT scan, what is the probability that he/she has the lung cancer?

Answer: We will solve this problem using the EvidenceBasedReasoning algorithm. Using Bayes' theorem, we already solved the problem and knew the correct answer for that question is

P(cancer | positive x-ray & positive CT scan) = 0.96

Here, we are going to show you that our new framework of reasoning will help us to get the result easier. The following is the analysis and steps of finding the answer:

First, we decide what is the question: after read the problem statement, we know the question is: P(cancer | positive x-ray & positive CT scan) = ?

Second, we calculate the degree for the prior probability (having cancer in a population) and the degrees for the two tests (x-ray and CT scan):

degree(prior)	= 10 $\log_{10}$ (0.002)	= - 27	(get from 2:998)
degree(x-ray)	= 10 $\log_{10}$ (14.17)	= 11.5	(get from 0.85/0.06)
degree(CT scan)	= 10 $\log_{10}$ (850)	= 29.3	(get from 0.85/0.001)

Third, we get the overall degree by adding all above degree values:

degree(answer) = 13.8

Fourth, we extract the answer in terms of probability by using Formula 7:

strength(answer) = $10^{\text{degree(answer)} / 10}$

= $10^{13.8 / 10}$

= 23.99

Convert to probability, it equals P = 23.99 / (23.99 + 1) = 0.96

Thus the final answer is:

P(cancer | positive x-ray & positive CT scan) = 0.96
(same value as the one got in Example 2)

As you can see, our evidence based reasoning is easier than the original Bayes' theorem in dealing with many evidences. One thing to point out is that our evidence based reasoning can be used in many areas. For example, in bioinformatics, data mining, category classification, etc., just to name a few.

7. Knowledge management in bio-information system architecture

We described the fundamentals of computer reasoning and proposed an EvidenceBasedReasoning algorithm. In this section, we will introduce a framework of knowledge management in the context of bio-information system architectures. Based on this framework, we will introduce a prototype implementation of the Bio-information knowledge management system.

7.1 Knowledge management framework

In a typical knowledge management system, there are many components. Figure 5 shows an information system architecture upon which we base our reasoning framework and knowledge management methods.

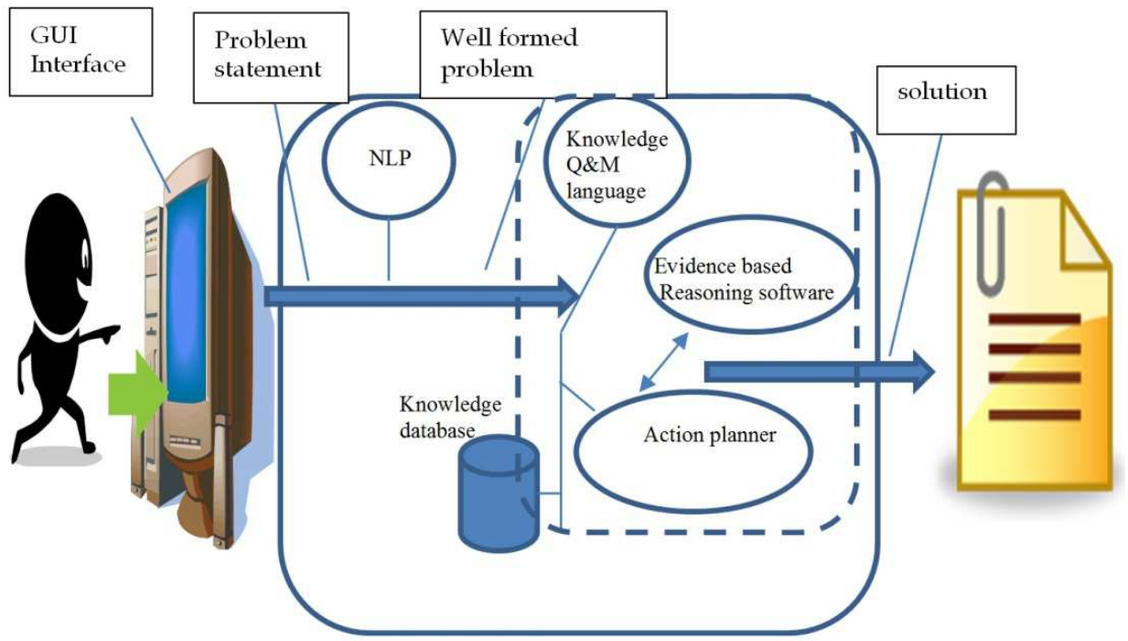


Fig. 5. A bio-information knowledge management framework

In Figure 5, the NLP stands for the natural language processor. NLP is used to translate a problem written in natural language such as English or Chinese into a well formed problem statement that is understood by reasoning engine which is enclosed inside the dotted region in Figure 5. The reasoning engine consists of Knowledge Query and Manipulation Language (KQML), the Evidence based reasoning software, and the Action planner. KQML is used to manage data stored in the knowledge database. Its role in an expert system is much like the role that the SQL language played in a database management system. Action planner is the component that drives the system. One thing to point out is that the reasoning engine works with the help of the knowledge database.

7.2 The evidence based reasoning software

As you can see from Figure 5, the complete system of a bio-information knowledge management system has both software and hardware. In this presentation, we will focus on the software side. In particular, we will focus on one software component: the evidence based reasoning software (expert Software). We will assume that other components are already implemented and working.

7.3 The potential areas of using the evidence based reasoning system

One of the application areas of our evidence reasoning system is the terminal patient consulting bio-information system. When a patient is diagnosed with terminal illness, his first reaction is disbelieving. Then, he probably will ask questions like: what is the prognosis such as how long he can live; what is its etiology such as the cause of the disease; and whether it is hereditary. These questions are usually being answered by doctors or nurses. Often, answers that a patient got are generic based on average cases. Also, because of tight

schedules of doctors and nurses, sometimes the patient is not able to get answers in a timely manner. Here, we will develop a prototype system that will answer most of the questions that a terminal patient will have. Also, the answers from our system will be tailored to individual patients. Ideally, our system should be able to relieve a lot of burdens from doctors and nurses.

7.4 The evidence based reasoning software design ideas

We are going to develop a prototype of the evidence based reasoning software component. In the following, we will outline our design ideas.

Main design ideas: we strive the following:

1. The component should have a Graphic User Interface (GUI) to facilitate the use of the system. Figure 6 is a screen capture of the user interface.
2. It should be interactive. Based on the information in the knowledge data base, it may ask patient questions.
3. The component should be developed in such a way that it can be used to query different terminal illnesses, in other words, it should be generic.
4. There should be default values for those fields that a user does not input specific information.
5. The knowledge database should be separated from this component for the benefit of less coupling.

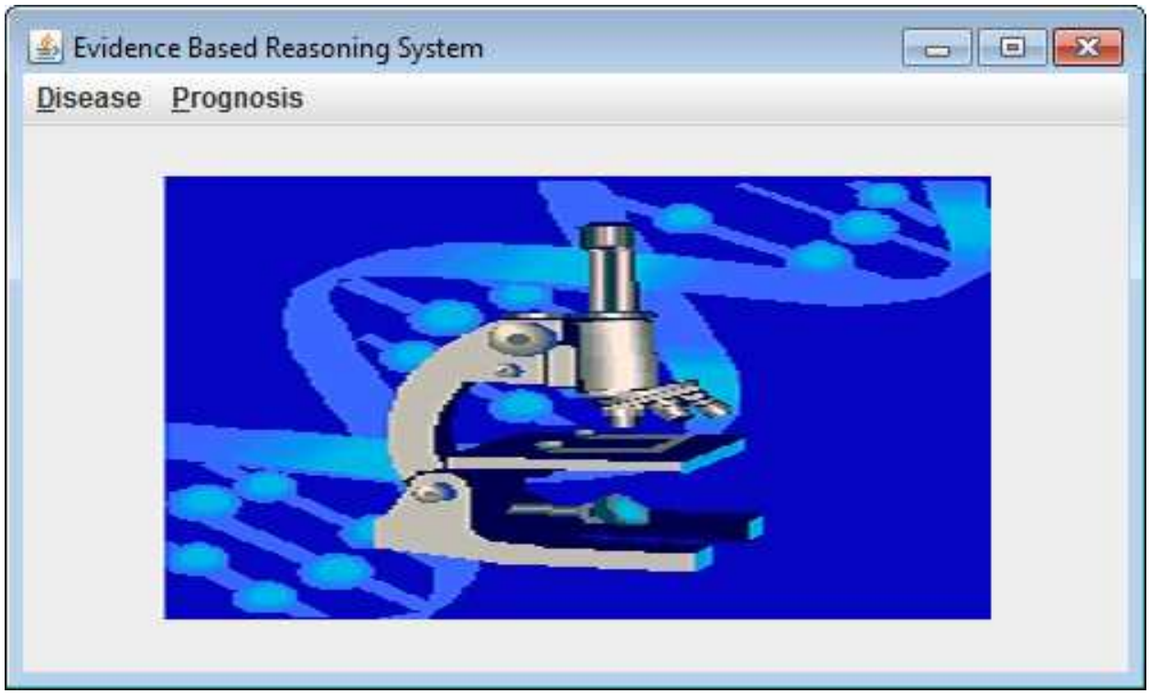


Fig. 6. A screen capture of the uesr interface for the evidence based reasoning system

With the above design ideas in mind, we will set the boundaries and make assumptions for the system. The following is the assumptions that we made.

Assumption: we assume the following:

1. All other components shown in Figure 5 are developed and working. The only component that we are focusing on is the evidence based reasoning software.
2. The diagnosis of the illness is already known.
3. The component will answer only predefined set of questions (most important to a patient) such as the cause of the disease (etiology), once diagnosed, how long a person can live (prognosis), etc.

To emphasize the main point, our implementation uses a simple design. Without losing generality, we loaded data from a file instead of asking the user to input them from a keyboard. We also watered down some features for the sake of simplicity. For example, the whole knowledge database is substituted by hard-coded logic.

7.5 A case example: Colorectal cancer

To help our presentation, we will use a medical case example to illustrate some features of our evidence based reasoning system. The medical case used is the colorectal cancer. And we will use the most common form of the colorectal cancer: the hereditary nonpolyposis colon cancer (HNPCC). This form of cancer is also called *Lynch syndrome*. The following is some facts related to this disease:

Some facts of colorectal cancer: "Cancer of the large bowel is second only to lung cancer as a cause of cancer death in the United States; 146,940 new cases occurred in 2004, and 56,730 deaths were due to colorectal cancer." (Kasper, 2005, p. 527) This disease has hereditary factors. "As many as 25% of patients with colorectal cancer have a family history of the disease, suggesting a hereditary predisposition." (Kasper, 2005, p. 527) Once diagnosed, the prognosis "is related to the depth of tumor penetration into the bowel wall and the presence of both regional lymph node involvement and distant metastases. These variables are incorporated into the staging system introduced by Dukes and applied to a TNM classification method, in which T represents the depth of tumor penetration, N the presence of lymph node involvement, and M the presence or absence of distant metastases (Table 1).

Stage				Approximate 5-yr survival, %
Dukes	TNM	Numerical	Pathologic Description	
A	T1N0M0	I	Cancer limited to mucosa and submucosa	>90
B_1	T2N0M0	I	Cancer extends into muscularis	85
B_2	T3N0M0	II	Cancer extends into or through serosa	70-80
C	TxN1M0	III	Cancer involves regional lymph nodes	35-65
D	TxNxM1	IV	Distant metastases (i.e., liver, lung)	5

Table 1. Staging of and Prognosis for Colorectal Cancer (Kasper, 2005, p. 529-530)

The prevalent belief of the cause of the disease is the interplay between the environment and the cancer suppressing genes. The reason why we have colorectal cancers (in fact, any type

of cancers) is because our body lost control to the cell growth. For normal cells, their growth is controlled by the information in their DNA. These cells know when to stop. On the other hand, for a cancer cell (either caused by spontaneous mutation or by hereditary predisposition), this control is lost. Thus, it will grow unchecked and with misshape. Environment factors such as high animal fat diet, radiation exposure, Streptococcus bacterial infection (bacteremia), inflammatory bowel disease, etc. make a person susceptible to colorectal cancers. But these factors do not mean a person has cancer. Cells in our body have innate ability to fight cancers. This ability is rested on the fact that normal cells have cancer suppressing genes. For example, "the long arm of chromosome 5 (including the APC gene)" is responsible for the suppression of one type of colon cancer (polyposis coli) development. "The loss of this genetic material (i.e., allelic loss) results in the absence of tumor-suppressor genes whose protein products would normally inhibit neoplastic growth." (Kasper, 2005, p. 528) Thus, when we see a cancer, it is the result of both the presence of the environmental risk factors and the absence of the cancer fighting genes.

7.6 Sample runs of the evidence based reasoning software

In this section, we will apply our prototype reasoning software to the case example introduced in the previous section. To show the effect of evidence, we will show two outputs: one with specific personal information and one without. The case information for the one that has no specific personal information is the following:

Case 1: we use the following general information (with no specific personal information):

> Suppose that the patient (Michael Dodd) is diagnosed with (HNPCC) colon cancer stage III.
> The information stored in the knowledge database is contained in Table 1.

Using the input information in case 1, we will get the default 5-year survival chance. Figure 7 is the output screen capture for case 1.

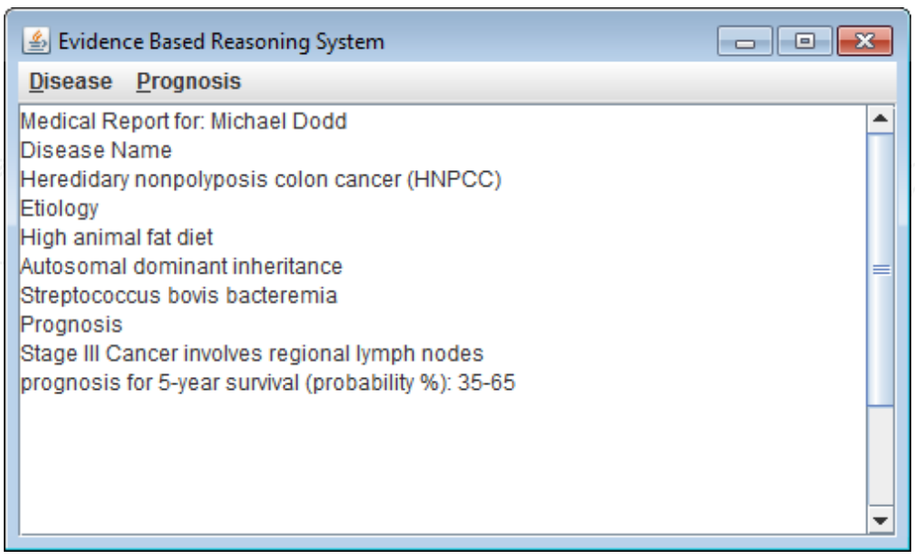


Fig. 7. A screen capture of the general 5-year survival probability for a person with colon cancer of stage III

The case information for the second run that has specific personal information is the following:

Case 2: we use the following specific information (with personal information):

Suppose that the patient (Michael Dodd) is diagnosed with (HNPCC) colon cancer stage III.
Michael's father had colon cancer, the time between the diagnosis and the death was 3 years.
Michael's older sister had colon cancer, the time between the diagnosis and the death was 4 years.
The information stored in the knowledge database is contained in Table 1.

Using the input information in case 2, we are able to get the revised 5-year survival chance. Figure 8 is the output screen capture for case 2.

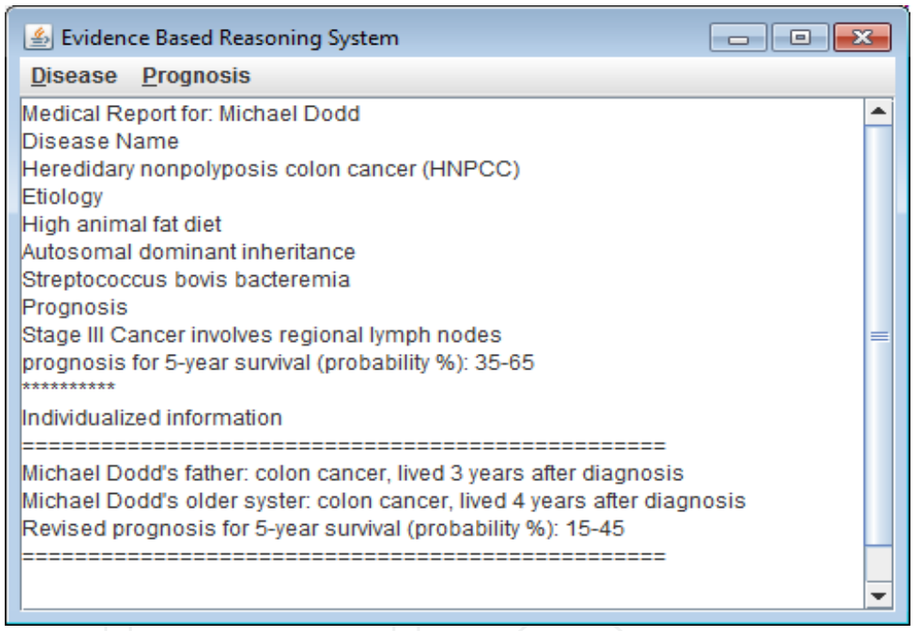


Fig. 8. A screen capture of the individualized 5-year survival probability for a person with colon cancer of stage III

As you can see from the output in Figure 8, the 5-year survival probability is revised down words. Since in this case, we have more information (patient's father's cancer history; patient's older sister's cancer history), the evidence based reasoning software takes the new information into account and produces more accurate output. With regard to the event of 5-year survival, these evidences reduce the probability. Thus, they are negative evidences according to our evidence theory. Specifically, the 5-year survival probability is revised from 35-65% down to 15-45%. The following is the rationale and steps to get this new result.

1. We first calculate the degree of prior probability (in this case, take the data from Table 1 (< 65%)) as follows:

$$\text{degree(prior)} = 10 \log_{10}(0.65) = -1.9$$

2. We rate the evidence E1 as follows: Michael's father lived 3 years after diagnosis as negative relative to the event of interesting: Michael is able to live 5 years. Considering Michael's father is his direct relative, we assign a degree(father's condition->Michael 5-year survival) = -1.
3. Similarly, we rate the evidence E2 as follows: Michael's older sister lived 4 years after diagnosis as negative relative to the event of interesting: Michael is able to live 5 years. Considering Michael's sister is his direct relative and the year 4 is pretty close to 5, we assign half degree(older sister's condition->Michael 5-year survival) = -.5.
4. Calculate the overall degree: FinalDegree = -1 + (-.5) + (-1.9) = -3.4.
5. Convert the degree to the final strength:

$$\text{strength(answer)} = 10^{\text{degree(answer)} / 10}$$

$$= 10^{-3.4 / 10}$$

$$= 1 / 2.2 = 0.45$$

6. Thus, the revised range will be: 15-45%.

Note: our assignments of degrees to the two evidences (in step 2, 3) are arbitrary in a sense that it is not verified. In real situation, we should determine these values by clinical trials.

As a consequence of these reasoning steps, the evidence reasoning software produces the revised survival probability as shown in Figure 8.

7.7 Java code for the reasoning software

The screen captures in the previous section are produced by Java code. We implemented the prototype using popular Java language. List 1 shows the code that produces the output screens.

List 1: Java code to produce the output screens

```
==========================================================================
import java.awt.event.*;
import java.awt.*;
import javax.swing.*;
import java.io.*;
import java.util.Scanner;

public class EvidenceBasedReasoningSystemApp extends JFrame implements ActionListener{
    private JMenuItem   diseaseMenuItem, closeMenuItem, prognosisMenuItem;
    private JTextArea contents;
    JLabel imageLabel;

    public static void main(String args[]) {
        EvidenceBasedReasoningSystemApp  frame = new EvidenceBasedReasoningSystemApp();
    }

    public EvidenceBasedReasoningSystemApp() {
        java.net.URL bkpic = getClass().getResource("genetics.jpg");
        ImageIcon bkpicture = new ImageIcon(bkpic);
```

```
        imageLabel = new JLabel (bkpicture);
        JMenuBar  menuBar = new JMenuBar();
        JMenu  diseaseMenu = new JMenu("Disease");
        diseaseMenuItem = new JMenuItem("Load Disease File");
        closeMenuItem = new JMenuItem("Close");
        diseaseMenu.add(diseaseMenuItem);
        diseaseMenu.add(closeMenuItem);
        JMenu  prognosisMenu = new JMenu("Prognosis");
        prognosisMenuItem = new JMenuItem("Display Prognosis");
        prognosisMenu.add(prognosisMenuItem);
        diseaseMenu.setMnemonic('D');
        prognosisMenu.setMnemonic('P');
        setJMenuBar(menuBar); //add menu bar to current frame
        menuBar.add(diseaseMenu);
        menuBar.add(prognosisMenu);
        add(imageLabel, BorderLayout.CENTER); //add background image
        contents = new JTextArea(20, 40);
        diseaseMenuItem.addActionListener(this); //subscribe events
        prognosisMenuItem.addActionListener(this);
        closeMenuItem.addActionListener(this);
        setSize(500, 300); //set the size of the frame
        setTitle("Evidence Based Reasoning System");
        setVisible(true);
        addWindowListener(new WindowAdapter()
        {
                public void windowClosing(WindowEvent event)
                        {shutDown();}
                }
        );
}

public void actionPerformed(ActionEvent e) {
        Object sourceObject = e.getSource();
        String string, tmp;

        if (sourceObject == diseaseMenuItem) {
                JFileChooser fileChooser = new JFileChooser(System.getProperty("user.dir"));
                String lineSeparator = System.getProperty("line.separator");
                JScrollPane scrollPane = new JScrollPane(contents);
                int result = fileChooser.showOpenDialog(this);

                if (result == JFileChooser.APPROVE_OPTION) {
                        File file = fileChooser.getSelectedFile();
                        try {
                                Scanner fileScan = new Scanner(file);
                                contents.setText("Medical Report for: ");
                                string = fileScan.nextLine();
                                contents.append(string + lineSeparator);
                                while (fileScan.hasNext()) {
                                        string = fileScan.nextLine();
                                        Scanner fieldScan = new Scanner(string);
                                        fieldScan.useDelimiter("/");
                                        while (fieldScan.hasNext()) {
```

```
                                tmp = fieldScan.next();
                                contents.append(tmp + lineSeparator);
                            }
                        }
                        imageLabel.setVisible(false);
                        add(scrollPane, BorderLayout.CENTER);
                    } catch(IOException ioe) {
                        ioe.printStackTrace();
                        return;
                    }
                }
            }
            else if (sourceObject == prognosisMenuItem) {
                JFileChooser fileChooser = new JFileChooser(System.getProperty("user.dir"));
                String lineSeparator = System.getProperty("line.separator");
                JScrollPane scrollPane = new JScrollPane(contents);
                int result = fileChooser.showOpenDialog(this);

                if (result == JFileChooser.APPROVE_OPTION) {
                    File file = fileChooser.getSelectedFile();
                    try {
                        Scanner fileScan = new Scanner(file);
                        contents.setText("Medical Report for: ");
                        string = fileScan.nextLine();
                        contents.append(string + lineSeparator);
                        while (fileScan.hasNext()) {
                            string = fileScan.nextLine();
                            Scanner fieldScan = new Scanner(string);
                            fieldScan.useDelimiter("/");
                            while (fieldScan.hasNext()) {
                                tmp = fieldScan.next();
                                contents.append(tmp + lineSeparator);
                            }
                        }
                        imageLabel.setVisible(false);
                        add(scrollPane, BorderLayout.CENTER);
                    } catch(IOException ioe) {
                        ioe.printStackTrace();
                        return;
                    }
                }
            }
            else if (sourceObject == closeMenuItem) {
                shutDown();
            }
        }

        public void shutDown() {
            System.exit(0);
        }
}
```

==========

8. Conclusion

In this chapter, we described the relationships among data, knowledge, and intelligence. We proposed one reasoning theory: evidence based reasoning theory. We gave the Java code for the implementation of a prototype. The future work includes more detailed mapping between the evidence strength value and its percentage change; the implementation of missing components such as the knowledge database, the beef up of the watered down features.

9. Acknowledgement

I want to give thanks to my family for their support for this book writing project: Enlu Peng, Yuqing Peng, and Daniel Jian.

10. References

Asy'arie, A., & Pribadi, A. (2009). Automatic News Articles Classification in Indonesian Language by Using Naive Bayes Classifier method, *In Proceedings of iiWAS2009*, Kuala Lumpur, Malaysia, 2009

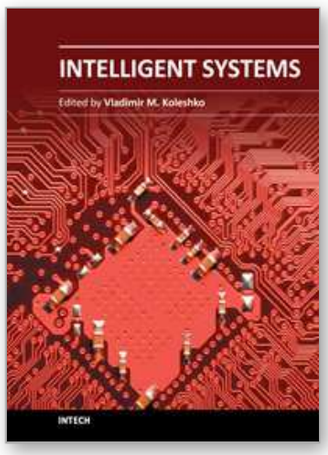
Bellinger, G. (2004). Knowledge Management—Emerging Perspectives, (Internet resouce: http://www.systems-thinking.org/kmgmt/kmgmt.htm. Retrieved on 5/30/2011)

Fujita, H.; and et al. (2010). Virtual Doctor System (VDS): Medical Decision Reasoning Based On Physical and Mental Ontologies, *In Proceedings of IEA/AIE'10 Proceedings of the 23rd international conference on Industrial engineering and other applications of applied intelligent systems*, Volume Part III, 2010

Jin, X.; and et al. (2007). Automatic Web Pages Categorization with ReliefF and Hidden Naive Bayes, *In Proceedings of SAC 2007*, pp. 617-621, Seoul, Korea, March, 2007

Kasper, D.; and et al. (2005). *Harrison's Principles of Internal Medicine* (Sixteenth Edition), McGraw-Hill Companies, Inc., ISBN 0-07-139140-1, USA

Williams, L. & Hopper, P. (2003). *Understaning Medical Surgical Nursing* (2nd Edition), F. A. Davis Company, ISBN 0-8036-1037-8, Philadelphia, PA, USA



Intelligent Systems

Edited by Prof. Vladimir M. Koleshko

ISBN 978-953-51-0054-6
Hard cover, 366 pages
Publisher InTech
Published online 02, March, 2012
Published in print edition March, 2012

This book is dedicated to intelligent systems of broad-spectrum application, such as personal and social biosafety or use of intelligent sensory micro-nanosystems such as "e-nose", "e-tongue" and "e-eye". In addition to that, effective acquiring information, knowledge management and improved knowledge transfer in any media, as well as modeling its information content using meta-and hyper heuristics and semantic reasoning all benefit from the systems covered in this book. Intelligent systems can also be applied in education and generating the intelligent distributed eLearning architecture, as well as in a large number of technical fields, such as industrial design, manufacturing and utilization, e.g., in precision agriculture, cartography, electric power distribution systems, intelligent building management systems, drilling operations etc. Furthermore, decision making using fuzzy logic models, computational recognition of comprehension uncertainty and the joint synthesis of goals and means of intelligent behavior biosystems, as well as diagnostic and human support in the healthcare environment have also been made easier.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Kuodi Jian (2012). Knowledge Management in Bio-Information Systems, Intelligent Systems, Prof. Vladimir M. Koleshko (Ed.), ISBN: 978-953-51-0054-6, InTech, Available from: http://www.intechopen.com/books/intelligent-systems/knowledge-management-in-bio-information-systems

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen