New Generation Computing, 17(1999)349-358
Ohmsha, Ltd. and Springer-Verlag

NEW GENERATION COMPUTING

Intelligent Systems Using Web-pages as Knowledge Base for Statistical Decision Making

Kazunori FUJIMOTO and Kazumitsu MATSUZAWA
NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto
6190237 JAPAN
{fujimoto,matuzawa}@cslab.kecl.ntt.co.jp

Received 1 February 1999
Revised manuscript received 26 April 1999

Abstract In this paper, we propose an approach to the construction of an intelligent system that handles various domain information provided on the Internet. The intelligent system adopts statistical decision-making as its reasoning framework and automatically constructs probabilistic knowledge, required for its decision-making, from Web-pages. This construction of probabilistic knowledge is carried out using a *probability interpretation* idea that transforms statements in Web-pages into constraints on the subjective probabilities of a person who describes the statements. In this paper, we particularly focus on describing the basic idea of our approach and on discussing difficulties in our approach, including our perspective.

Keywords: Decision Support Systems, Internet Users, Statistical Decision Making, Knowledge Acquisition, Probabilistic Reasoning.

§1 Introduction

The Internet has now become widespread and this has enabled humans to immediately acquire information from all over the world. Correspondingly, a lot of techniques are being researched to gather valuable information from the Internet and to use it for various human judgments, e.g., AAAI symposiums.[8,10] Although a great deal of effort has been made on developing such techniques, judgment using information on the Internet is still difficult. The amount of acquired information may be large because it is gathered from enormous texts on the Internet. Moreover, it may not be well-known because the information on the Internet is frequently updated. These characteristics of information on the Internet do not allow humans to quickly use it for their judgment. In order to resolve such difficulties, we have researched the construction of an intelligent system that handles such information instead of humans and derives a conclusion for humans as advice.

First, we will describe our approach with a concrete example where a person chooses a type of a product, e.g., a digital camera, to suit his or her preference. When a person chooses a product, he or she first has to know the product types put on the market. Such information can be easily acquired by using the Internet and the person can learn a lot of names of different alternatives. The person will next try to choose a type from the alternatives on the condition that it is the best one from a sense of his or her preference, e.g., the camera has excellent image quality, good portability, and so on. Such a choice is often based on specifications that are gathered from the Internet together with the name of various types. However, the specification documents may contain unknown technical terms for the person, or may contain up-to-date specifications whose effect is not obvious to the person. The goal of our research is to develop an intelligent system that provides the best choices for the person by handling such information instead of the person. We call this system as DSIU system, which gives a Decision Support for Internet Users. Fig. 1 shows the information flow of a DSIU system.

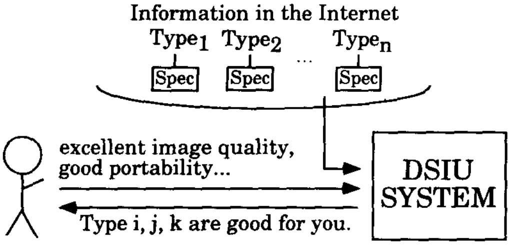
Information in the Internet
Type_1
Type_2
Type_n
Spec
Spec
Spec
excellent image quality, good portability...
Type i, j, k are good for you.
DSIU SYSTEM

Fig. 1 Information Flow of DSIU System

Here we have extracted two important capabilities for a reasoning framework of DSIU systems as below.

- Handling a lack of information (e.g., the lack of a specification of each type of a product) : information required for reasoning is not always on the Internet. Even if it is available, all of the information cannot always be gathered within a limited time. DSIU systems should give advice on these conditions to a person.
- Handling each person's sense of values (e.g., person's weight of preference to each property of a product) : advice should be provided to give the most profit to each different person. To evaluate such profits, DSIU systems should take account of each person's sense of values.

In order for DSIU systems to have these capabilities, the framework of the statistical decision-making[2] is adopted as their reasoning framework. This framework can represent uncertainty caused by a lack of information as "probabilities", and represent a person's sense of value as "utilities", respectively. A prospect of profits acquired by each choice is quantified as "expected utility values", which are calculated from probabilities and utilities. In terms of the example, an expected utility value for a choice implies a prospect of the fitness of properties of the chosen type for the person's preference of the properties. As a result, a type taking the largest expected utility value can be provided as the best type for the person. To adopt this framework, a method has to be developed for acquiring probabilistic knowledge as probabilities and utility knowledge as utilities. In this paper, we propose DSIU systems that particularly focus on knowledge acquisition methods to acquire knowledge for statistical decision-making framework. Section 2 describes our approach to acquiring knowledge base for DSIU systems. Section 2.1 explains that probabilistic knowledge is more difficult to acquire than utility knowledge in constructing a knowledge base. It also describes our approach using Web-pages for acquiring probabilistic knowledge. Section 2.2 describes an idea of *probability interpretation* that can acquire probabilistic knowledge from Web-pages. Section 3 illustrates a process of reasoning in DSIU systems with a concrete example. Section 4 discusses difficulties in our approach, including our perspective.

§2 Our Approach to Knowledge Acquisition

2.1 Acquisition of Probabilities and Utilities

In order to make a statistical decision, we have to acquire: (1) probabilities that enables us to learn probabilistic prospects of the result generated by the decision, and (2) utilities that enables us to numerically learn one's profits received from the result. The second condition is restricted with respect to one's interest, e.g., the image quality and portability; although it may require a significant effort, it can be acquired using various conventional techniques.[16] On the other hand, the first condition has to correspond to information gathered from the Internet, namely, information containing various up-to-date content. From this view point, one may say that this is not an actual approach such that humans prepare probabilistic knowledge with their hands because it will take enormous costs to preserve probabilistic knowledge as large and up-to-date. Thus, the problem "How can we acquire probabilistic knowledge that corresponds to information on the Internet?" is one of the most important in constructing a knowledge base in DSIU systems.

To cope with this problem, we turn our attention to the fact that Web-pages contain some information that is available for the construction of probabilistic knowledge. For example, the Internet provides various articles on various types of a product, e.g., advertisements, performance reports, and so on. It may contain a sentence describing a relationship between specifications, e.g., "the lens l of this camera improves image quality". By extracting such relationship, we may acquire dependencies between lens l and image quality, which is available for the construction of probabilistic knowledge. In addition to this, such statements on the Internet contain various content and are frequently updated themselves. Our approach is to automatically construct probabilistic knowledge from Web-page statements*[1] and to preserve the knowledge to correspond to information gathered from the Internet.

As just mentioned, the essence of our DSIU systems can be formalized here.

1. Utility function U is acquired from a system user, and probability function Pr is acquired from Web-pages.
2. A alternative T_i that makes expected utility $E_{(\mathrm{Pr},\mathrm{U})}(T_i)$ maximum, namely, T_i that achieves
$$\max_{T_i} E_{(\mathrm{Pr},\mathrm{U})}(T_i)$$
is chosen from a set of alternatives $\{T_1, ..., T_n\}$ to the system user.
3. This alternative T_i is provided to the system user.

2.2 Probability Interpretation

To construct probabilistic knowledge from statements gathered from the Internet, the statements have to be expressed as probabilities of a domain. Statements on the Internet may contain information of subjective probabilities*[2] of those humans who describe the statements. For example, a sentence:

"Specification s of this camera improves image quality."

may be interpreted as an argument "adopting s makes the possibility of the camera having high image quality more likely" of the person who describes the original sentence. This argument is formalized as subjective probabilities of the person:

$$\Pr(Q = high|Spec = s) > \Pr(Q = high),$$

where Q and *Spec* denote discrete random variables representing image quality and specifications, respectively. We call this transformation from statements into constraints on the subjective probabilities of a person who describes the statements *probability interpretation*. This probability interpretation transforms statements on the Internet into constraints on probabilities in a domain. As a result, probabilistic knowledge can be constructed from gathered constraints on probabilities by using various conventional reasoning methods.[5]

Much research has been done to acquire subjective probabilities with respect to their suitable representations[7,11] and respect to their relationships to verbal expressions.[3,4] What seems to be lacking, however, is researches on relationships between subjective probabilities and various information on the Internet. We aim for the extraction of subjective probabilities from information provided on the Internet. Therefore, our research has two new important issues: (1) which information on the Internet can be available for subjective probabilities, and (2) which forms of subjective probabilities are suitable for representing information on the Internet. For these issues, we have to extend conventional

*[1] This is not restricted to Web-pages. We particularly focus on them because of the convenience of html tags for text processing.

*[2] This term "subjective probabilities" is used to represent personal probabilities that reflect human knowledge, and distinguish them from physical probabilities.

techniques for acquiring subjective probabilities to techniques that properly handle information on the Internet.

Pearl is one researcher who studies the probabilistic interpretation of meanings in logical sentences.[12] Goldszmidt extended the idea to take account of linguistic quantifiers, e.g., believable, unlikely, in each sentence.[6] To apply such ideas to information on the Internet, various kinds of meanings that are not restricted to verbal meanings should be considered. For example, the location of each sentence in a text and the font sizes for each sentence may contain useful information for constructing probabilistic knowledge. As a concrete example, we use an article containing two sentences "spec s_1 improves image quality" and "spec s_2 improves image quality". When these sentences are independently described, no difference between spec s_1 and s_2 can be acquired. However, they are quite different when the font size of the former one is larger than the latter. In this case, we can guess that the person who describes the sentences would like to emphasize the former sentence over the latter one because a larger font size is used for the former sentence. As a result, this interpretation gives us a kind of knowledge so that spec s_1 is much more effective in improving image quality than spec s_2. This knowledge can be formalized with a parameter α determined by the ratio of their font sizes as

$$\Pr(Q = high|Spec = s_1) > \Pr(Q = high|Spec = s_2) + \alpha,$$

where α is a positive real number less than 1. As previously mentioned, in order to acquire subjective probabilities from information in the Internet, probability interpretation takes account not only of verbal meanings in each sentence but also of meanings in text structures like locations, font sizes, and so on.

The process of probability interpretation is illustrated below using a communication model of two persons A and B (Fig. 2). In Fig. 2, person A sends a statement to the Internet. Person B receives the statement and then interprets it as information of the subjective probabilities of person A. This interpretation is based on the subjectivity of person B. As a result, person B acquires person A's subjective probabilities containing the subjectivity of person B. The main idea of probability interpretation is that interpretation rules are constructed by using the subjectivity of the receiving person. By using these interpretation rules, statements on the Internet are transformed into constraints on probabilities in a domain. Thus, probability interpretation can acquire probability constraints that are available to the construction of probabilistic knowledge.

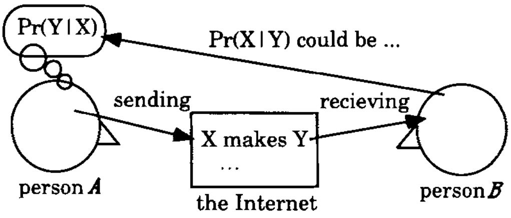

Fig. 2 Model of Probability Interpretation

§3 Example

This section describes a process of statistical decision making in DSIU systems with a concrete example. As an example, we use a decision problem that chooses the best selection for a system user from two types of digital cameras, denoted type A and B, respectively. Four discrete random variables $Q \in \{q_1, q_2\}$, $M \in \{m_1, m_2\}$, $F \in \{f_1, f_2\}$,and $L \in \{l_1, l_2\}$ represents possible attributes of image quality, portability, focusing method, and lens type respectively. The specifications of type A and B are "$F = f_1$ and $L = l_1$" and "$F = f_2$ and $L = l_2$", respectively. The three constraints on probabilities are described below.

$$\Pr(Q = q_1 | F = f_1) > \Pr(Q = q_1) \tag{1}$$

$$\Pr(Q = q_1 | L = l_1) > Pr(Q = q_1 | L = l_2) \tag{2}$$

$$\Pr(M = m_1 | F = f_2) > Pr(M = m_1 | F = f_1) \tag{3}$$

Equations (1), (1), and (2) mean "types adopting f_1 have a high quality", "types adopting l_1 have much higher image quality than types adopting l_2", and "types adopting f_2 have much more portability than types adopting f_1", respectively. These kinds of statements often appear in an advertisement provided on the Internet. Four posterior probabilities $\Pr(Q = q_1|\text{type } A)$, $\Pr(M = m_1|\text{type } A)$, $\Pr(Q = q_1|\text{type } B)$, and $\Pr(M = m_1|\text{type } B)$ are calculated on the above conditions. A derivation method[5)] proposed by Druzdzel was used.[*3] These probabilities are shown in Table 1.

Table 1 Posterior Probabilities

	$\Pr(Q = q_1 \mid \text{type})$	$\Pr(M = m_1 \mid \text{type})$
type A	0.68	0.38
type B	0.32	0.62

The expected utility value for each choice can be calculated using the posterior probabilities in Table 1. The value for choosing type A is $0.68\alpha_Q + 0.38\alpha_M$ and the value for B is $0.32\alpha_Q + 0.62\alpha_M$, where α_Q and α_M take a positive real number and denote a utility value for variables Q and M. These utility values α_Q and α_M are acquired from the system user to reflect his or her preferences of the image quality and portability of digital cameras. A type with larger expected utility values is better type for the system user because it means that the possibility of the type suiting the system user's preference is much higher. According to this principle, a type taking the largest expected utility value can be provided as the best type for the system user. In this example, when the value for type A is larger, namely $3\alpha_Q \geq 2\alpha_M$, type A is chosen. This means type A, which has higher image quality, is recommended to system users who prefer image quality over portability. When the value for type B is larger, namely $3\alpha_Q < 2\alpha_M$, type B is chosen. This means type B, which has better portability, is recommended to system users who prefer portability over image

[*3] Druzdzel proposed a method for deriving second-order probability distributions from a set of constraints on probabilities. In this section, the posterior probabilities are calculated as the expected value of each second-order probability distribution.

quality. As mentioned above, a DSIU system provides advice to system users by using probability constraints acquired from the Internet.

§4 Discussion

To apply the idea of probability interpretation, semantic relationships between words in Web-pages have to be analyzed. However, such analyses can be difficult to do with conventional natural language processing techniques. Therefore, we have begun to examine new acquisition methods that focus on customs humans commonly use in describing information. For example, advertisements are pieces of information that have a clear aim: that of advertising something. In advertisements, writers usually describe significant features in the beginning of an advertisement in order to catch readers' attention. In terms of digital cameras, the feature "1.4 million pixels" represents a superior CCD specification for image quality. Thus, advertisements for cameras that have high CCDs usually include the CCD description in the beginning of the advertisement. Conversely, advertisements for cameras that have inferior CCDs, e.g., 0.35 million, usually put the CCD description near the end of the advertisement and put descriptions of other features, e.g., being light weight, at the beginning. Based on this custom, we hypothesized that we could identify the rank of a particular specification by the location of each spec-description in an advertisement. To examine the validity of this hypothesis, we randomly selected 60 digital cameras and gathered advertisements on them from home-pages provided by their makers. Then, focusing on CCD pixel sizes, we divided the cameras into three categories: less than 0.8 million, between 0.8 and 1.3 million, and above 1.3 million. We then examined the locations of the CCD descriptions in the advertisements. Table 2 shows the results of this examination.

Table 2 Location of the CCD Description in Each Advertisement

CCD pixel sizes	Number of advertisements	Average line number of first mention
– 0.8 million	10	21.6
0.8 – 1.3 million	18	15.2
1.3 – million	32	4.0

In Table 2, the second column shows the number of advertisements gathered that correspond to each camera category. The third column shows the average line number on which CCD is first mentioned in an advertisement.[*4] In Table 2, we can see that for advertisements that correspond to larger pixel CCDs, line numbers tend to be smaller, namely, the first description of CCD appears relatively beginning of the advertisement. We have implemented an acquisition method that is based on this rating custom and have confirmed it is highly accurate when applied to ranking several specifications in digital cameras. As mentioned in Section 2.2, such customs are not only limited to location of descriptions in a Web-page, but also apply to font sizes for sentences, the number of times words

[*4] The lines that mentioned CCD were detected automatically as lines that included the word "CCD" or its synonyms.

are repeated, and so on. By combining these customs with conventional natural language processing techniques, the idea of probability interpretation can be applied to statements on the Internet. It is true that such automated acquisition methods do give some incorrect knowledge to DSIU systems mainly because of analytical failures in text processing. As a result, DSIU systems may sometimes provide wrong choices. However, as long as they mainly provide correct choices, DSIU systems are still useful tools for Internet users. This is similar to Internet search engines, which do not always provide the best URLs but are still useful tools for Internet users.

The process of constructing a knowledge base in a DSIU system may be computationally intractable when the size of domain knowledge become large. It may also be required to handle inconsistency in a knowledge base because statements on the Internet may be conflicting. Therefore, an efficient computation method and an inconsistency handling method have to be introduced to realize DSIU systems. Many conventional techniques provided in Artificial Intelligence can be used to develop these techniques. For example, a technique for dividing a knowledge base into a small subset of knowledge can be used to decrease computational complexity.[9)] Belief maintenance systems[14)] that identify a consistent set of probabilistic knowledge within an inconsistent one is available for handling inconsistency. The difficulties in computation and in inconsistency will be handled properly by introducing these conventional techniques and some approximation techniques to the DSIU system.

On the other hand, the research field of Knowledge Discovery[13)] has recently been attractive for extracting or organizing knowledge in large databases. Data mining techniques are used for extraction in this field. The method of knowledge construction described in this paper may be considered as a technique that corresponds to the data mining techniques. Most data mining techniques consider data as a "case" and extract knowledge by focusing on the frequency of the appearance of the case.[1,15)] In contrast, our approach considers data as a constraint on human knowledge and extracts knowledge to reconstruct human knowledge. In this viewpoint, we consider our approach as a communication based approach for acquiring knowledge, as illustrated in Fig. 2. When statements are directly described by a human, it is also important to research such statements not as a case but as a clue for reconstructing human knowledge.

§5 Conclusion

In this paper, we proposed an approach to the construction of an intelligent system, called DSIU system, that handles information on the Internet. Information provided on the Internet contains various domains' contents and is frequently updated. In order to construct a DSIU system to handle such information, the knowledge base in the system has to be maintain various up-to-date knowledge that corresponds to information on the Internet. We turned our attention to the fact that such knowledge exists on the Internet itself, and proposed an idea of *probability interpretation* that enables the transformation of statements on the Internet into probabilistic knowledge. There are some difficulties when constructing a DSIU system using this idea of probability interpretation. Two main difficulties were discussed in detail in terms of extracting knowledge from a text and in the construction of knowledge base within a limited time. A perspective that resolves these difficulties was described by referring to an extension of conventional techniques provided in Artificial Intelligence. As a result, we believe DSIU systems will be realized by adopting the approach described in this paper. We have started to examine the idea of probability interpretation with advertisements of digital cameras provided on the Internet. This result of the examination will be reported sometime soon.

Acknowledgements

The authors would like to thank Professor Norio Okino, University of Shiga Prefecture, for guidance and advice in this study. The authors also thank Hideto Kazawa, NTT Communication Science Laboratories, for providing statistical data on the Internet.

References

1) Agrawal, R., Imielinski, T. and Swami, A., "Database Mining: A Performance Perspective," *IEEE Transactions on Knowledge and Data Engineering*, *5*, *6*, pp. 914–925, 1993.

2) Berger, J. O., *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, 1985.

3) Beyth-Marom, R., "How Probable is Probable? A Numerical Translation of Verbal Probability Expressions," *Journal of Forecasting*, *1*, pp. 257–269, 1982.

4) Druzdzel, M., "Verbal Uncertainty Expressions: Literature Review," *CMU-EPP*-1990-03-02, Carnegie Mellon University, 1989.

5) Druzdzel, M. and van der Gaag, L., "Elicitation of Probabilities for Belief Networks: Combining Qualitative and Quantitative Information," in *Proc. of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence* (UAI-95), Morgan Kaufmann, pp. 141–148, 1995.

6) Goldszmidt, M. and Pearl, J., "Qualitative Probabilities for Default Reasoning, Belief Revision, and Causal Modeling," *Artificial Intelligence*, *84*, *1*, pp. 57–112, 1996.

7) Heckerman, D. and Jimison, H., "A Bayesian Perspective on Confidence," in *Uncertainty in Artificial Intelligence 3* (Kanal, L. N., Levitt, T. S., and Lemmer, J. F. ed.), Elsevier Science Publisher, pp. 149–160, 1989.

8) Knoblock, C. and Levy, A. ed., Information Gathering from Heterogeneous, Distributed Environments, *AAAI Spring Symposium Series Technical Reports*, The AAAI Press, 1995.

9) Lauritzen, S. L. and Spiegelhalter, D. J., "Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems," *Journal of the Royal Statistical Society*, *50*, *2*, pp. 157–224, 1988.

10) Mahech, K. ed., Natural Language Proc. for the World Wide Web, *AAAI Spring Symposium Series Technical Reports*, The AAAI Press, 1997.

11) Paaß, G., "Second Order Probabilities for Uncertain and Conflicting Evidence," in *Uncertainty in Artificial Intelligence 6* (Bonissone, P. P., Henrion, M., Kanal, L. N. and Lemmer, J. F. ed.), Elsevier Science Publisher, pp. 447–456, 1991.

12) Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 1988.

13) Piatetsky-Shapiro, G., "Data Mining and Knowledge Discovery References," URL: http://www.kdnuggets.com/references.html.

14) Ramoni, M. and Riva, A., "Belief Maintenance in Bayesian Networks," in *Proc. of the Tenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–94)*, Morgan Kaufmann, pp. 498–505, 1994.

15) Ramoni, M. and Sebastiani, P., "Learning Bayesian Networks from Incomplete Databases," in *Proc. of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–97)*, Morgan Kaufmann, pp. 401–408, 1997.

16) Winterfeldt, D. V. and Edwards, W., *Decision Analysis and Behavioral Research*, Cambridge University Press, 1986.

Kazunori Fujimoto: He received bachelor's degree from Department of Electrical Engineering, Doshisha University, Japan, in 1989, and master's degree from Division of Applied Systems Science, Kyoto University, Japan, in 1992. From there, he joined NTT Electrical Communications Laboratories, Tokyo, Japan, and has been engaged in research on Artificial Intelligence. He is currently interested in probabilistic reasoning, knowledge acquisition, and especially in quantitative approaches to research in human cognition and behavior. Mr. Fujimoto is a member of Decision Analysis Society, The Behaviormetric Society of Japan, Japanese Society for Artificial Intelligence, Information Processing Society of Japan, and Japanese Society for Fuzzy Theory and Systems.

Kazumitsu Matsuzawa: He received B.S. and M.S. degrees in electronic engineering from Tokyo Institute of Technology, Tokyo, Japan, in 1975 and 1977. From there, he joined NTT Electrical Communications Laboratories, Tokyo, Japan, and has been engaged in research on computer architecture and the design of LSI. He is currently concerned with AI technology. Mr. Matsuzawa is a member of The Institute of Electronics, Information and Communication Engineers, Information Processing Society of Japan, Japanese Society for Artificial Intelligence, and Japanese Society for Fuzzy Theory and Systems.