A CBR Based Prediction Method for Web Aquatic Products Prices

Hongchun Yuan [1] **Ying Chen** [2] **Jinling Ju** [1]

[1]College of Information Technology, Shanghai Fisheries University, Shanghai 200090, P. R. China
[2]School of Information Systems, University of Tasmania, Private Bag 87 Hobart, Australia

Abstract

The ability to scientifically forecast the price of aquatic products plays an important role in the healthy and sustainable development of aquaculture. This paper presents a method for forecasting aquatic product prices using Case-Based Reasoning. Some key processes include automatic extraction of web data, attribute-oriented induction based on concept hierarchy, generation and representation of cases, cases matching and similarity calculation, case evaluation and revision. An application has been implemented using these processes. Experiments have shown that the system can automatically extract web data from multiple websites with information on aquatic product prices and can effectively analyze and forecast prices accordingly.

Keywords: Aquatic products prices, Forecasting, Attribute-oriented induction, Case-based reasoning

1. Introduction

Case Based Reasoning (CBR) started with Schank and Abelson's work in 1977 [1]. In 1982 Schank presented the cognitive model of CBR in the book “Dynamic Memory” [2], and developed a CBR application system. It has been the basis of many CBR systems so far. After 30 years of development, CBR has been widely applied in many domains. For example, Ye Shi-ren proposed a multi-strategies center fisheries prediction method based on CBR [3]; Zhang Zhi-yan and others combined CBR reasoning mechanism with load forecasting and developed the electric load forecasting system [4]; Yu Yi-xin constructed a forecasting system based on CBR using fuzzy neural network to solve the short-term nodes load forecasting of the Medium Voltage Distribution Network [5]; Feng Jian-feng established an intelligent HABs (Harmful Algal Blooms) early-warning system based on CBR [6]; Sun Sheng-Tao and others introduced a credit risk evaluation system of financial organization based on CBR and RBR [7]; Canadian Abidi and others presented a personalized health information generation and delivery system [8]; Iranian Amani presented a case-based reasoning method for alarm filtering and correlation in telecommunication networks [9]. Currently, combining diversified technology with CBR to resolve more practical issues is the hotspot of CBR research.

Aquatic product price, as an information element of the aquatic product market, is the specific embodiment of the supply and demand of aquatic products. It is crucial to use scientific forecasts to provide the basis for decisions on the main aquatic products market levels [10]. Scholars in China have begun research on forecasting the aquatic product prices. For example, Yuan yong-ming discussed the construction of the basic forecasting system design together with its key components and used a linear dynamic model as forecast examples to describe the extensive application and feasibility of the system [11]; Zhang Xiao-shuan predicted farmer price of aquatic products in 2000 via a time series decomposed model sampled between 1978 and 1999. He also constructed the domain ontology model of the aquatic products forecasting [12]. However, forecasting aquatic product prices by using the year as granularity failed to meet the demand for the market subject at all levels. In order to predict future trends of aquatic product prices with varied time granularity, the authors present a web-based application that can collect and analyze current irregular publication of web information on aquatic product prices using a web data automatic extraction algorithm. The application also applies the principles from CBR and concept hierarchy for some key processes in aquatic product price forecast. Results of multiple experiments for testing the application show that both the automatic extraction algorithm of web data and the aquatic product price forecasts are effective and feasible.

2. Basic concepts and principles

Both concept hierarchy and CBR have been applied as the basis for the price forecast method proposed in the paper.

2.1. The concept hierarchy

Definition 1: Concept hierarchy is a kind of ordinal set T=(H, $\prec$), H is a limited concept set, $\prec$ is a kind of ordinal.

Definition 2: If x, y∈H, x $\prec$ y, x≠y, and no concept z(z ∈H), make x$\prec$z and z$\prec$y, then Concept y is called a nearest ancestor of concept x.

Definition 3: If there exists a maximum element and a set H_l (l =0,1,…,(n-1)) in H of concept hierarchy T =(H , $\prec$), then the resulting formula is

$$H = \bigcup_{l=0}^{n-1} H_l, and\, H_i \cap H_j = \phi \qquad (i \neq j) \qquad (1)$$

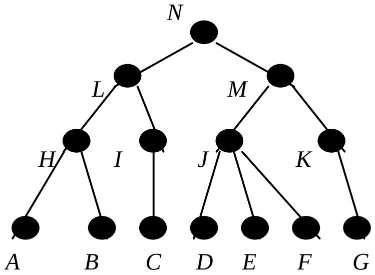

Fig. 1: An example of concept hierarchy.

An example of positive concept hierarchy is shown in Fig.1. Its maximum element is N, and H0={N}, H1={L, M}, H2={H, I, J, K}, and H3={A, B, C, D, E, F, G}. (Concept hierarchies mentioned in later parts are all positive concept hierarchies.)

2.2. CBR principles

The basic idea of CBR is to solve current problems by studying and revising the successful solutions of the past similar problems. While the system is used to solve a problem, it compares the new problem with the deposited cases in the system, and retrieves the case that is similar to the new problem. The solution to the new problem can be obtained by adopting and revising the method or the model solving similar cases. Once the current problem is solved, the description of the problem, the solving solution and the final outcome can be stored in a CBR system as a revised case in order to solve the next similar problem.

There are four basic steps in a typical CBR process: retrieve, reuse, revise and retain. The procedure is illustrated in Fig. 2.

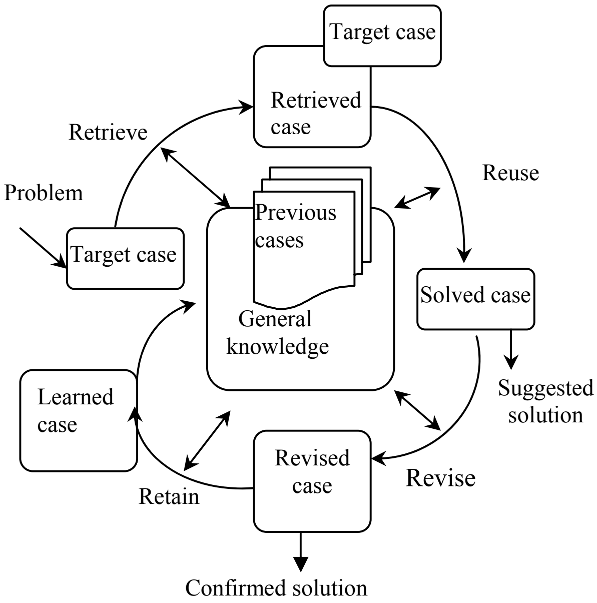

Fig. 2: CBR principles

3. Proposed forecast method and key processes

The proposed method contains a CBR-based flowchart for price forecast and five key processes.

3.1. Method of aquatic products prices forecast

The flowchart of CBR-based aquatic product price forecast is shown in Fig. 3. The process consists of two main parts, namely, case generation and case use. Case generation follows the procedure like this: firstly it uses the automatic web data extraction algorithm to automatically obtain aquatic product price data from various websites and stores them to a database; secondly, according to need, it filters special aquatic product price information that spreads over the years from the database and uses it to form subject data; thirdly it generates the data with granularity ranging from a day to ten days or a month by using the attribute-oriented induction based on the concept hierarchy; lastly the data will be organized into a case and then added into the case base of the subject. The process of case use follows this procedure: firstly, according to the demand of time granularity, input recent aquatic product price information to form a target case; secondly, through case retrieval, a case set that is similar to the target case is retrieved from the subject case base followed by a calculation of the similarity between the target case and each case in the case base; thirdly, based on the level of similarity after the calculation, one case from the case base is selected which, after some revise, is obtained as a new case; lastly the output price trends of the revised new case is used for forecast and retained also in the subject price case base.

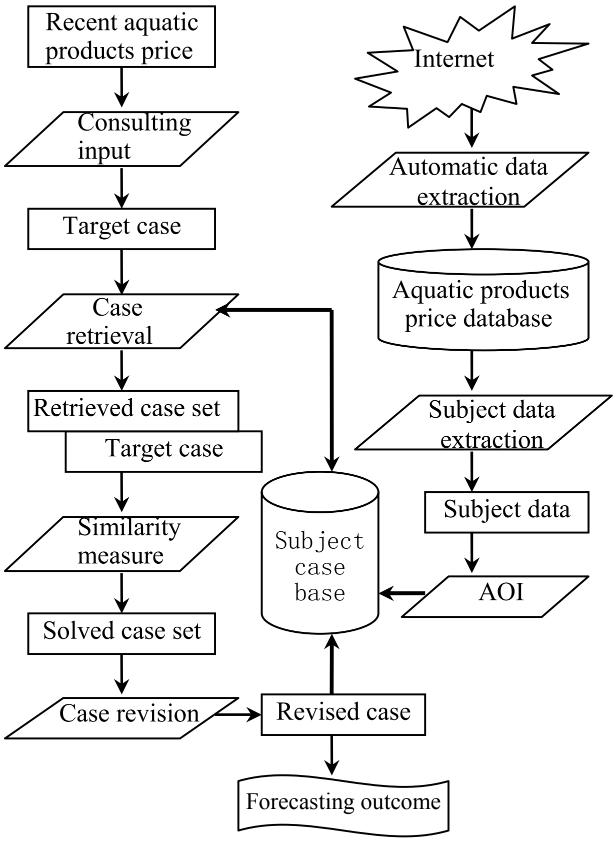

Fig. 3: Flowchart of CBR-based aquatic products price forecast.

3.2. Key processes

Five key processes have contributed towards the forecast of aquatic product prices.

3.2.1. Algorithms of Automatic Data Extraction

Keywords Distance Algorithm (KDA) is used to automatically extract data (See Fig.4). Keywords could represent subjects that users are interested in. According to the probability distribution of the locations of the keywords, KDA can locate an Interested Block (IB), which often contains just one layer of <Table> and </Table> structure. Once a table is located, it is easy to identify the tuples of the data according to <Tr> and </Tr> tags, and then the columns according to <Td> and </Td> tags. The steps of the KDA are presented in the following diagram.

3.2.2. Attribute-Oriented Induction Algorithm

```
INPUT：A set of keywords $W = \{w_1, w_2, \ldots, w_k\}$,
        Normalized Document (ND)
OUTPUT：Interested Block (IB)
BEGIN
Step1: Counting the frequency $n_i$ and a position set
        $P_i = \{p_{i1}, p_{i2}, \ldots, p_{in}\}$ of each keyword $w_i$ in
        the ND, let $P = \sum_i^{\|W\|} \bigcup p_i$, $\|W\|$ represents
        the number of the keywords.
Step2: For each keyword $w_i$, if $n_i = 1$, then the
        table block B where the keywords are
        located is the interested block IB, return
        IB.
Step3: Sort ( P )
Step4: Scanning P, If two adjacent items $p_l$ and
        $p_{l+1}$ come from the same keyword set $P_i$,
        then delete $p_l$.
Step5: Span =$\|W\|$, for the set P' that is the result
        set of P after Step 4, scan it with a
        distance of Span, and get the Manhattan
        distance d with those Span items.
Step6: According to the ascending order of
      previous distances, Detect whether the
      elements in each span locate in the same tag
      couple of <Table> and </Table> that can be
      called scope B, If true then return IB=B;
        Else check the next Span.
END
```

Fig. 4: Keywords Distance Algorithm（KDA）.

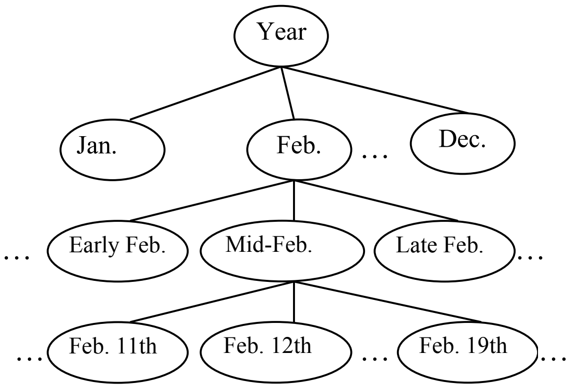

Fig. 5: Time granularity concept hierarchy.

Attribute-oriented induction is a more effective method of data generalization. Its use in price forecasts for aquatic products need background knowledge that the time granularity should be expressed in the form of concept hierarchy. Time granularity concept hierarchy is shown in Fig. 5. Attribute-oriented induction algorithm is shown in Fig. 6.

```
INPUT : (1) Subject data, (2) Concept hierarchy
        of time granularity
OUTPUT : The concentrated subject data of
specified level
BEGIN
Step1:  Designate the level that the subject data
        should be generated among the concept
        hierarchy of time granularity;
Step2:  Up along the concept hierarchy of time
        granularity, replace the date in subject data
        table with the designated time granularity
        value in step 1;(For example: replace the
        date "February 11th of certain year" with
        "Mid-February of certain year")
Step3:  Merge the records with the same time
        granularity value in the subject data table,
        take the average price and then the
        concentrated subject data is generated.
END
```

Fig. 6: Attribute-oriented induction algorithm.

3.2.3. Generation and expression of cases

According to the characteristics of aquatic product price forecasts, the generation and expression of a dynamic case base is used. This method uses a Database Management System (DBMS), which has powerful functions to manipulate data, to organize, store and manage the basic data of aquatic product prices. When inputting a target case, the data retrieval and generalization according to the supplied price data on special time granularity are carried out and a dynamic case base is then generated. For example, knowing the lobster's lowest average price is 396.67 Chinese yuan in September 2006 and 370.00 Chinese yuan in October respectively in Shanghai Tongchuan Aquatic Products Market, to forecast the lobster's lowest average price in November 2006, the data in the aquatic product prices base can be filtered and categorized according to the subject and a dynamic case base can be generated as shown in Table 1.

3.2.4. Case matching and similarity calculation

In order to achieve the matching between the target case and each case in the dynamic case base, the Euclidean Distance method is used to calculate similarity for the price growth rate. Assuming the target case Y has price data of m months (or ten days), then the similarity between case X_i and target case Y can be defined as in Formula (2).

$$Sim(X_i, Y) = 1 - \{\sum_{h=2}^{m}(v_{i,k} - w_k)^2\}^{1/2} \qquad (2)$$

Where, $v_{i,k}$ is the price growth rate of the *k*th month of the *i*th case, w_k is the price growth rate of the *k*th month of the target case Y.

Year	Month	The lowest average price (￥)
…	…	…
2004	9	370.00
	10	273.33
	11	280.38
2005	9	373.05
	10	323.33
	11	294.50
…	…	…

Table 1: a dynamic case base.

3.2.5. Evaluation and revision of cases

The main basis for evaluation and revision of cases is their similarity. The idea is to set a threshold d of the similarity according to experience, and the cases which meet Formula (3) below are as solved cases.

$$Sim(X_i, Y) \geq d \qquad (3)$$

For all the solved cases that with similarity values greater than the threshold, calculate the weighted average of the price growth rate according to the similarity, and then work out the price growth rate of the target case in the forecasting month (or other time granularity), as illustrated Formula (4).

$$v' = \frac{\sum_{i=1}^{n} Sim(X_i, Y)(v_{i,k+1})}{\sum_{i=1}^{n} Sim(X_i, Y)} \qquad (4)$$

Where n is the number of solved cases, $v_{i,k+1}$ is the price growth rate of the *i*th solved case in the forecasting month (or other time granularity), v' is the price growth rate of the target case.

At last, the forecasting result of the target case can be calculated by Formula (5).

$$p' = p * (1 + v') \qquad (5)$$

Where, *P* is the price of the target case in the month (or other time granularity) before the forecast one.

4. Experiment and Analysis

Based on the forecasting methods above, a corresponding software tool is realized and is tested with aquatic product price data issued by a aquatic products consulting website (http://www.china-fisheries.org/). Table 2 shows the forecasting results of lowest average price of lobster of 500-1000 grams by using the above methods.

Time	Forecast results （￥）	Actual price （￥）	Relative error
May 2007	412.67	395.00	4.47%
Nov. 2006	337.01	344.17	-2.08%
Dec. 2006	354.47	330.00	7.42%

Table 2: The lobster's price forecast results in Shanghai Tongchuan Market.

An analysis of the experiment suggests the following findings:

(1) An automatic data extraction tool based on KDA can quickly and efficiently obtain the aquatic product price information from the designated aquatic websites, and store them to the local database.

(2) Notwithstanding that an aquatic product consulting website issues a large number of price information, the price information for a specific market especially a specific aquatic product, is still insufficient. Besides, the time interval is irregular. Very often fine time granularity data is absent. As a result, it becomes very challenging to accurately forecast prices on the corresponding concept level.

(3) The forecasting method introduced here adopts attribute-oriented induction based on the concept hierarchy. By improving fine time granularity of data to rough time, it can solve the problem caused by irregular publication of price information or fine granularity data loss.

(4) When CBR is adopted on the rough time granularity level of the concept hierarchy, it is feasible to calculate the similarity for aquatic product prices growth rate or whether the similarity value of the weighted average of the solved cases is greater than the threshold.

5. Conclusions

With the extensive application of internet technology in the field of aquaculture, using a website to release aquatic product price has become a trend. The paper presents a CBR-based approach for price forecasting of aquatic products and constructs a forecasting system which can automatically collect Internet data on aquatic product prices. When inputting a target case, the system extracts and filters data from an existing data base to produce a dynamic case base according to a user's demand regarding the product type and time granularity. It then calculates the similarity between the target case and cases in the dynamic case base. The cases with similarity higher than the similarity threshold will be taken as solved cases that are weighted average to form a revised case. The revised case is used to predict the outcome of the target case. Results of multiple experiments have indicated that the proposed method can effectively analyze and forecast aquatic product prices using source data from dedicated web sites.

Acknowledgement

This work is partially supported by Shanghai Education Committee (Grant No. 06KZ016) and National "Eleventh Five Years" Science and Technology Support Program of China (Grant No. 2006BAD10A11).

References

[1] R. Schank and R.R. Abelson, Goals and Understanding. *Erlbanum: Eksevier Science*, 1977.

[2] R. Schank, DynamicMemory: A Theory of Reminding and Learning in Computers and People. *London: Cambridge University Press*, 1982.

[3] S.R. Ye and Z.Z. Shi, Predicting Center Fishery Based on CBR. *High Technology Letters*, 5:64-68, 2001.

[4] Z.Y. Zhang, J.B. Chen and J. Li, An Expert System of Short Term Load Forecasting Based on Sample Inference. *Northeast Electric Power Technology*, 7:35-37,2005.

[5] Y.X. Yu and J.Z. Wu, CBRFNN-Based Short-term Nodal Load Forecasting for Middle Voltage Distribution Networks. *Proceedings of the CSEE* , 12:18-23,2005.

[6] J.F. Feng, Y. Qu, H.M. Li, F. Shen and H.L. Wang, Research on Intelligence HABs Early-Warning System Based On CBR. *Ocean Technology*, 2:63-65,2005.

[7] S.T. Sun and X.S, Wang, The Realization and Discussion of the Credit Risk Evaluation System. *Journal of Communication and Computer*, 5:119-125, 2006.

[8] S.S.R. Abidi, A Case Base Reasoning Framework to Author Personalized Health Maintenance Information. *Proceedings of the 15 the IEEE Symposium on Computer-Based Medical Systems* (CBMS 2002), Maribor, Slovenia, 2006.

[9] N. Amani, M. Fathi and M. Dehghan, A Case-Based Reasoning Method For Alarm Filtering And Correlation In Telecommunication Networks. *Canadian Conference on Electrical and Computer Engineering*, pp.2182-2186, 2005.

[10] X.S. Zhang, J. Zhang, Z.T. Fu and W.S. Mu, A Time Series Decomposed Model for Forecasting Farmer Price of Aquatic Product. *Journal of Agriculture Mechanization Research*, 2:86-88, 2004.

[11] Y.M. Yuan, H.Y. Zhang and Y.H. He, Online Price Forecasting System on Aquatic Product. *Chinese Fisheries Economics*, 6:30-32, 2005.

[12] X.S. Zhang, J. Zhang and Z.T. Fu, Aquatic Products Price Forecasting Support System Based on Agent. *Computer Engineering*, 8:65-67,2004.