FLOREON$^+$: using Case-Based Reasoning in a system for flood prediction

T. Kocyan[1], J. Martinovic[1], J. Unucka[2] & I. Vondrak[1]

[1] Faculty of Electrical Engineering and Computer Science, VSB – TU Ostrava, Czech Republic

[2] Faculty of Mining and Geology, VSB – TU Ostrava, Czech Republic

Abstract

Human consumption and other natural factors are changing our climate and the world we live in. The changing climate and the increased risk of flooding around the world are among some of today's most urgent social issues. There are many resources and technology available to assist in damage reduction and minimizing consequences related to this natural phenomenon. This is possible because the development of information technology has expanded into many branches of human life. This article describes the development of a knowledge system created to assist in limiting cases of natural disasters. Our goal is to design and create a system to be used by either professional and non-professional organizations, or individuals, as a tool for exchanging knowledge. Our Disaster Prevention system (DIP) combines data mining and the analysis of past events to predict future natural disasters. We apply the Case-Based Reasoning (CBR) method. The principle of this method is based on collecting data (knowledge, experiences, etc.) and then applying this information to achieve new solutions. Tests have shown that by using information about past events, it is possible to determine future patterns of nature (e.g. weather conditions on specific land surfaces). These patterns can be used for describing the risk of future natural disasters and enables us to deduce the threat imposed by them. Analyzing this data and creating new solutions may assist in the fight to minimize damage incurred by these disasters.

Keywords: Case-Based Reasoning, disaster management, case-based prediction, information retrieval.

WIT Transactions on The Built Environment, Vol 110, © 2009 WIT Press
www.witpress.com, ISSN 1743-3509 (on-line)
doi:10.2495/DMAN090241

1 Introduction

Human consumption and other natural factors are changing our climate and the world we live in. Current issues include global climate change and the imminent risk of flooding. Today, the development of information technology, and its mass use in various sectors, has resulted in significant advances. We now have technology and facilities that can prevent damage, or at least minimize the consequences which may arise by these phenomena.

Unlike most existing systems, our disaster prevention system (DIP) does not apply a strict mathematical description of the different variables. It applies Case-Based Reasoning methodology (CBR) [10], which is the principle of collecting data (experience) and its subsequent application for deriving new solutions. Any such "experience" in the terminology of CBR is called a *Case*. This is a set of appropriately characterized attributes from a particular situation in the past [1].

The user specifies the current situation – location and threat of a specific natural phenomenon. The system then searches for similar cases in the past related to the given situation and determines an approach for dealing with this situation. Based on the similarity of these cases with user-specified situations, DIP derives the appropriate measures and eventual threats. These results are then presented to the user.

DIP is currently prepared for manual data entry and related measures (http://www.floreon.eu). We are now preparing to automatically input selected information. The interface is designed to be connected to any source respecting the structure of data.

2 System architecture

An essential element of the structure is characterized as a vertical cross-section of the entire architecture which consists of these six layers: The *user* enters a *natural phenomenon* (from a particular *territory* that has caused some *damage* or *impairments*) into the system. The way the situation is resolved then becomes more *reliable*.

Based on this collected data, the system is capable of deriving solutions for new cases via CBR methodology. The accuracy of estimation will be in direct proportion to the increase in the number of cases already correctly entered into the system.

2.1 Phenomena

One of the key elements of the system is natural phenomena. Based on a specific phenomena's strength power in given territory, the system is able to search existing cases of similar situations from which potential damages and consequences may be predicted. We have decided that the phenomena in the system will be divided into a hierarchy and stored in the database in order to develop a system modularly and independent of the phenomena. This structure will clarify individual categories making them easier to work with.

Based on consultations with experts, we created a tree of categories. To distinguish between the strength of individual phenomena, we have divided each into several degrees of intensity. These degrees of intensity were determined based on consultations with experts or with the aid of tested methods for gauging the strength of a phenomenon.

A weighted vector is a part of every end category of phenomena and identifies the importance of individual aspects when comparing two cases. Each event includes three weighted vectors that form a hierarchical structure. At the micro level, there is a vector for the surface according to the specifications of Corine Land Cover [4]. Seven components with a strong emphasis on the specification of the surface are in the middle part, namely:

influence of the location of areas of interest, influence of Corine Land Cover as a complex, influence of river networks, influence of the slope, influence of the orientation, influence of altitude, influence of neighboring territories.

The highest weighted vector (or vector for the whole case) then determines: weight of strength significance, weight of phenomenon duration, weight for similarity (on which the phenomenon operates).

2.2 Territory

Exposure of phenomena to a certain type and composition may naturally have the effect of natural and environmental disasters. As already mentioned, this system will serve to prevent and mitigate these disasters. The aim of this section is to find an efficient method for describing a territory.

The territory is separated into two parts. The first part (raster), determines the composition of a given area or, more specifically: the percentage of individual components of the surface of a given territory. The second part (vector), simply describes the river network, which plays a major role in the operation of any element.

The Corine Land Cover method, which offers various types of surfaces and their locations, is used to describe the territory.

The approximate structure of water is characterized by two indicators – nodes and flows.

Nodes: nodes are places where the river branches, runs, or breaks at an angle exceeding the limit angle.

Flows: flows are used as vectors starting in FNODE and orientation to TNODE. FNODE and TNODE are labels of individual nodes.

The territory described by a user will be expressed as a vector containing the specific parameters for a given area (the vector describing the territory), and will be presented in the following format:

Area location				Corine Land Cover Data								River net												Area slope (°)	Area orientation (°)	DTM (m)
				Area of interest				Surrounding areas				Area of interest						Surrounding areas								
X_0	Y_0	X_1	Y_1	C_0	C_1	··	C_N	C_0	C_1	··	C_N	P	O	S	R	%R	%V	P	O	S	R	%R	%V			

Figure 1: Data structure of the vector.

The first part of the vector is location, to be determined using the S-JTSK system. The second part of the vector is created by using Corine data, which is made up of individual components of Corine Land Cover, and expressed in percentages (or values in the range of 0-1). These values are identified for the particular field of interest as well as its surroundings. This may affect the territory in which there has been a phenomenon. The third part is the vector river network. As is the case with Corine data, the river network consists of two parts – the area of interest and its surroundings.

River inflow and outflow, river junctions, river branching, and river segments are all searched in a given territory. The slope, orientation, and digital terrain model is used for a more detailed description of the field. These properties constitute the last three attributes of our vector.

2.3 Solutions, the consequences and damage

The solution is a set of measures and other actions bound to a specific phenomenon, leading to the minimizing of consequences or damage. The term "solution" is therefore used to present sandbags, for example, which are used to prevent the flow of water during floods. Since our goal was to create an intelligent system, we did not settle for a mere determination of whether or not a given solution was used in a specific case.

This is why we define the structure as illustrated in Figure 2.

The solution is composed of indicators assigning its jurisdiction to the phenomenon and mainly outlines potential values that may be assigned during this step. If we define the above example with sandbags as a solution (field values), we also determine the recommended height for stacking such bags.

The solutions defined as above are only general rules for a situation (the mere abstraction). To be able to convert this information into a ”tangible” form, we define a particular solution that represents the specific action in specific circumstances. The main media information is made up of attributes: the success of solution and index value. Since the minimum and maximum for each solution is defined, the following formula is sufficient: $RealValue = Solution_{\min} + RatioValue * (Solution_{\max} - Solution_{\min})$.

The value ratio was introduced in order to unify the scale of all phenomena, thus providing an improved overview. Maximum and minimum values may obviously provide inconsistent values, thus, the introduction of new, specific solutions to exceed the scale interval is automatically extended and its values are converted.

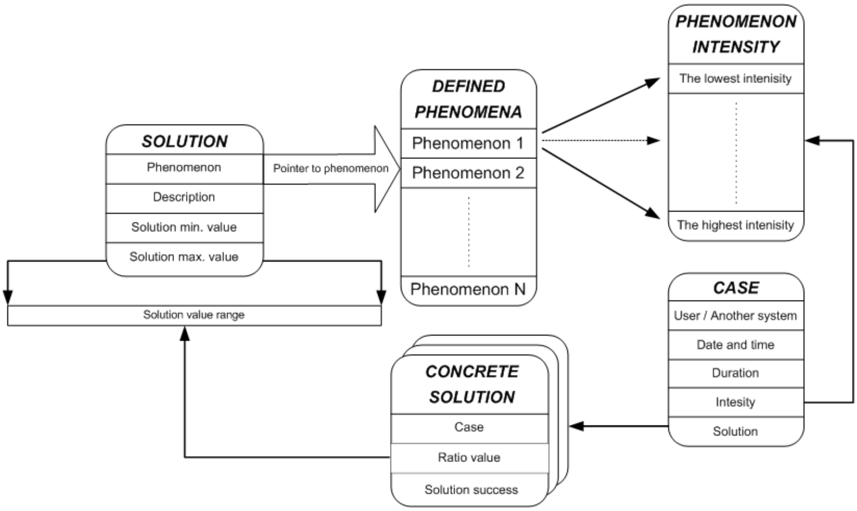


Figure 2: Block diagram of follow-up data on the subsequent derivation of a solution.

2.3.1 Consequences and damage

A formal definition of consequences and damages is no longer necessary. Now unwanted phenomena occurring after exposure to a natural element, despite all efforts and measures taken to prevent it, becomes the priority.

2.3.2 Derivation and solution

The system DIP which we develop is not only a warehouse of recorded cases of the past, but is able to intelligently on the basis of experience to derive a solution for the situation which currently threatens. Selecting the most appropriate action is carried out using the methodology of CBR and its course is shown in the picture.

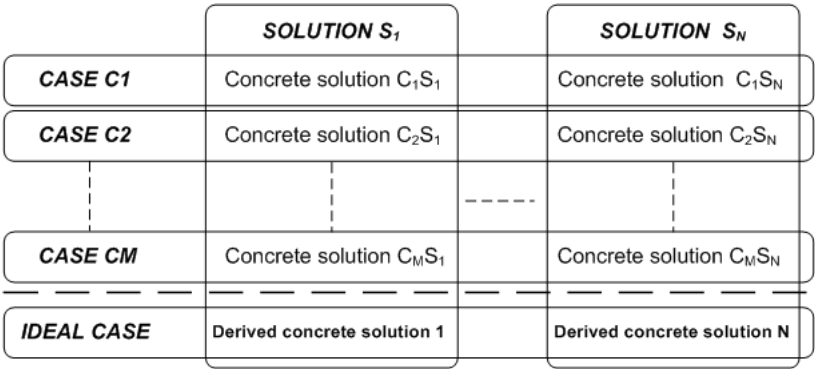
Activity (score) of the searched phenomenon can be summarized as follows:

1. The user chooses a list of phenomenon that directly threatens the field of action, intensity, estimated duration and other parameters, described in the scenarios above.
2. The system then monitors all cases, which relate to similar situations, with corresponding information describing the territory.
3. The system monitors all solutions used for the given phenomenon.
4. The deductible matrix is created and evaluates a list of appropriate consequences and damage.
5. The deductible matrix is created and evaluates a list of appropriate measures.

A detailed approach of individual items is described in the following paragraphs.

2.3.3 Observations of similarities

The first step for deriving an appropriate solution is the calculation of similarity [10] found in sample cases which the user entered. This number is expressed in



	SOLUTION S_1		SOLUTION S_N
CASE C1	Concrete solution C_1S_1		Concrete solution C_1S_N
CASE C2	Concrete solution C_2S_1		Concrete solution C_2S_N
⋮	⋮	-----	⋮
CASE CM	Concrete solution C_MS_1		Concrete solution C_MS_N
IDEAL CASE	Derived concrete solution 1		Derived concrete solution N

Figure 3: Appearance of deductive matrix.

the interval $CaseSimIndex \in \langle 0, 1\rangle$, where 0 indicates the maximum diversity and the number 1 symbolizes the identity. The similarity is now calculated on two levels as defined by the phenomenon and its weight vector, as shown in the figure.

The landscape and its surface structure are derived at the lowest level of similarity, creating the attribute $LandSimIndex \in \langle 0, 1\rangle$ unit scope. This value, combined with the duration and intensity of the phenomenon, determines the final congruity of the entire case $CaseSimIndex$.

$$CaseSimIndex = SimilarityVector \times WeightedVector, \qquad (1)$$

where *SimilarityVector* consists of the following members:

***LandSimIndex*:** similarity of the territory,
***WeightedVector*:** similarity of the length of exposure to the phenomenon.

An important part of the formula is *WeightedVector* where its components determine how important different parameters of vector similarity *SimilarityVector* are for the calculation. This enables us to easily determine which components can be ignored and which we need to highlight, thus speeding up the calculation.

2.3.4 Creation of deductive matrix

Once we have traced cases and evaluated their similarity, we can create a deductive matrix that will help us to derive specific solutions. The matrix has the following form:

Rows form our searched cases, while columns define all the solutions used for the phenomenon. The last line of the matrix presents the ideal case.

2.3.5 The calculation of the ideal values

Each particular solution is a system characterized by the proportionate value "success" which is defined as the number $SolutionSuccess \in \langle 0, \infty\rangle$. A value of less than 1 indicates that the solution was insufficient, whereas a higher index indicates an unnecessary waste of resources. Ideally calculated values then become the basis upon which the entire algorithm works – seeking solutions for crisis situations.

Derivation of the ideal solution is carried out as follows:

1. For each specific solution in the current column of a deductive matrix:
   (a) Calculate the ideal value and solution for the given event.
   (b) Add to the calculated value the potential impact of a phenomenon's duration.
   (c) Add to the calculated value the potential impact of exposure to a phenomenon.
   (d) The system derives a weight for the similarity of the earth's surface to surface similarities of both cases and the weight vector.
   (e) The system derives a calculation for the total weight of the final solution.
2. The system calculates the recommended weight average using collected data and adds the item to a list of recommended measures for the situation.

2.3.6 The calculation of probable consequences

Even after measures are evaluated, it is still necessary to alert the user of the consequences of this type of phenomenon and to what extent it is likely that this situation will affect the user. Derivation is equivalent to the calculation of the recommended solutions of a deductive matrix; the calculation is simplified by the fact that consequences are monitored only as a binary value (whether it happened or not). The result is the number of operations in the interval *AfterEffectProbability* $\in \langle 0.1 \rangle$, where we obtain the probability of % after we multiply the number by 100.

3 Testing

Derivation of the correct solution depends not only on previous cases, but on the content of a weighted vector, as well. In the following example, the model will be verified by the reaction of component weights at the highest level, namely the influence of intensity and duration of the phenomenon.

To demonstrate, we shall overlook the reliability factor of individual cases and automatically consider a case to be true. Then, we assume that the ideal value in all cases is equal to 100. We also limit solutions to that of one type; this same procedure can be repeated for an infinite number of types of solutions.

The algorithm works in such a way that an ideal solution is derived for each individual case. Consequently, the derived value of the final solution is collected from all the proceeding weighted average values and then returned to the user. Individual weights are then determined by an index of case similarity and then investigated.

The value of 0 in the weighting vector means that this item has no effect on the outcome; the value of 1 indicates a linear relationship. Table 1 was based on weight vector having all entries zero (the intensity equal to 0, the effect of duration effect equal 0). (A header of the following terms is used in the test sheet: *CASE* (serial number of the test case), *Intensity* (the intensity of exposure to the phenomenon),

Table 1: Derived values for the zero weight vector.

CASE	Intensity	Duration	Solution	Solution success	Ideal value	Derived value
1	0,2	100	100	1,0	100	100
2	0,4	100	50	0,5	100	100
...	...	...	...	...	...	...
9	0,8	200	930	9,3	100	100
10	1,0	200	1035	10,4	100	100
Case	0,2	100	—	—	—	100

Table 2: Derived values – emphasis on phenomenon intensity.

CASE	Intensity	Duration	Solution	Solution success	Ideal value	Derived value
1	0,2	100	100	1,0	100	100,0
2	0,4	100	50	0,5	100	50,0
...	...	...	...	...	...	...
9	0,8	200	930	9,3	100	25,0
10	1,0	200	1035	10,4	100	20,0
Case	0,2	100	—	—	—	45,7

Duration (duration of exposure to the phenomenon), *Solution* (the value which is typical for the current case), *Solution Success* (title, to what extent it was successful solution), *Ideal value* (value, which is ideal for the current case) and *Derived value* (the ideal value derived by the system for the current case).) The last row shows the case for which we are seeking a solution.

Table 1 it is clear that the reference value of all individual cases is equal to 100. This is because the intensity or duration has no effect on the derivation of a recommended value and remains as an ideal solution for the current case.

However, if we set the weight vector of the phenomenon so that the emphasis is on the ratio of intensity, the values change, as seen in the Table 2.

If the intensity of the phenomenon in the row is lower (or more precisely larger) than 0.2, the derived value in the line is larger (or more precisely lower). We assume that the resulting value will be much lower than the original 100. The result 45.6667 proves that the formula works correctly.

A similar experiment can be done for the length, if we assign the highest weight and vice versa overlook the effect of intensity. In the original table, there are length

Table 3: Derived values with an emphasis on the intensity of exposure to the phenomenon.

CASE	Intensity	Duration	Solution	Solution success	Ideal value	Derived value
1	0,2	100	100	1,0	100	100
2	0,4	100	50	0,5	100	100
...	...	...	...	...	...	...
9	0,8	200	930	9,3	100	50
10	1,0	200	1035	10,4	100	50
Case	0,2	100				75

Table 4: Derived values with an emphasis on the duration and intensity of exposure to the phenomenon.

CASE	Intensity	Duration	Solution	Solution success	Ideal value	Derived value
1	0,2	100	100	1,0	100	100,0
2	0,4	100	50	0,5	100	50,0
...	...	...	...	...	...	...
9	0,8	200	930	9,3	100	12,5
10	1,0	200	1035	10,4	100	10,0
Case	0,2	100				34,3

values of the same phenomena or longer, so again we expect to decrease the reference value. The resulting value of 75 3 satisfies the assumptions.

Finally, in order to be thorough, we need to present a case where there is an emphasis on both components. If the intensity and duration decreased the output value, we assume effect the same direction, but more intensively. The test result is shown in table 4.

The system obviously affects the output value according to the assumptions and provides the required functionality in this part.

3.1 The behavior of the system when entering an incorrect value

By proposing a method for users to indicate how successful their solutions were, we achieve a relatively high accuracy rate in calculating the recommended solution. On the other hand, this method places high demands on the accuracy of the

Table 5: Results distortion – different weights of abnormal case.

CASE	100%	50,00%	30,00%	10,00%
1	550,00	400,00	307,69	181,81
10	181,81	142,85	126,21	108,91
20	142,85	121,95	113,30	104,47
30	129,03	114,75	108,91	102,99
...	...	...	...	...
100	108,91	104,47	102,69	100,89
200	104,47	102,24	101,34	100,44
300	102,99	101,49	100,89	100,29
...	...	...	...	...
900	100,99	100,49	100,29	100,09
1000	100,89	100,44	100,26	100,08

estimate user. The objective of this test, therefore, is to map how the system preserves the introduction of errors (both unintentional and intentional), and how this value affects the calculation results for the other cases.

This testing data has been used in thousands of cases and is sufficient for this type of test. Furthermore, we chose different values for establishing credibility in points 10%, 30%, 50% and 100%. Simplified resulting data is illustrated in Table 5.

Thanks to our system of evaluating credibility, abnormal events with a low credibility rating should not overly influence the value, and with an increasing number of cases the effects should even be negligible.

The declining credibility of cases in which users decide for you, the effect on the output value decreases drastically.

With the decrease of cases where users make their own credibility decisions, the effect on the output value decreases drastically.

4 Conclusion

By designing and implementing the DIP system, we attempted to create a system to that, with a high level of abstraction, was able to monitor a large number of aspects in order to be applied in various sectors of crisis management. DIP will initially be deployed as part of the FLOREON$^+$ project [7, 9], which will be its main source of automated data. It will also be open to the general public, where users will be able to browse details of crisis situations and in this way it will serve as a potential manual for emergency situations.

An important architectural property of our system, upon which we have based the emphasis of our research and development, is its ability to provide solutions

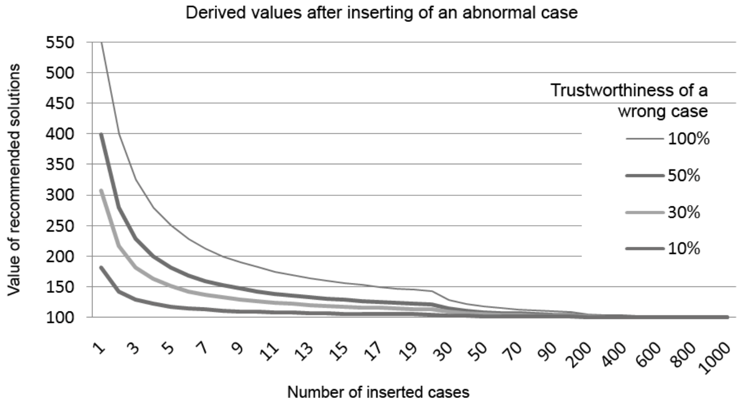


Figure 4: Graph of bias after the introduction of abnormalities.

regardless of influential factors. Therefore, our system is not limited to natural disasters (e.g. flooding, landslides, etc.). We can also apply this method in analyzing various situations such as automobile accidents and so on.

References

[1] Aamodt A., *Case-Based Reasoning: Foundational Issues*, Methodological Variations & System Approaches, AI Communications, Vol. 7 Nr. 1, March 1994.

[2] Bedient, P.B., Huber, W.C. & Vieux, B.C., *Hydrology and floodplain analysis*, 4th edition, Prentice Hall, London, 795 p., 2007.

[3] Bjarne K. Hanse & D. Riordan, *Fuzzy Case-Based Prediction of Cloud Ceiling and Visibility*, Dalhousie University, Halifax, NS, Canada, 2003.

[4] Bossard M., Feranec J. & Otahel J., *CORINE land cover technical guide*, Addendum 2000.

[5] Ishioka T., *Evaluation of Criteria for Information Retrieval*, The National Center for university Entrance Examinations, Japan, 2003.

[6] Jonov, M., Unucka, J. & Zidek, D., *The comparison of two floods in the Ole catchment - the possibilities of hydrological forecasting with the use of radar products*, Fifth European Conference on Radar in Meteorology and Hydrology ERAD 2008, Helsinki Finland, 2008.

[7] Martinovic, S., Stolfa, J., Kozusznik, J., Unucka, J. & Vondrak, I., *Floreon – the system for an emergent flood prediction*, In ECEC-FUBUTEC- EUROMEDIA, Porto, April 2008.

[8] Sun Z., Finnie G. & Weber K., *Integration of abductive CBR and deductive CBR*, Sch. of Inf. Technol., Bond Univ., Gold Coast, Qld., 2001.

[9] Vondrak, I., Martinovic, J., Kozusznik, J., Unucka, J. & Stolfa, S., *FLOREON - the System for an Emergent Flood Prediction*, 22nd EUROPEAN Conference on Modelling and Simulation ECMS 2008, Nicosia Cyprus, 2008.

[10] Watson I., *Applying Case-Based Reasoning: Techniques for Enterprise Systems*, Morgan Kaufman, 1997.