Cross-cultural and Environmental Data Analysis in Data Mining Processes for a Global Resilient Society

Yasushi KIYOKI[a,1], Xing CHEN[b], Anneli HEIMBÜRGER[c],
Petchporn CHAWAKITCHAREON[d], Virach SORNLERTLAMVANICH[e],
[a]*Graduate School of Media and Governance, Keio University SFC, Japan*
[b]*Kanagawa Institute of Technology, Japan*
[c]*University of Jyväskylä, Faculty of Information Technology, Finland*
[d]*Environmental Engineering Department, Faculty of Engineering, Chulalongkorn University, Thailand*
[e]*Sirindhorn International Institute of Technology (SIIT), Thammasat University, Thailand*

Abstract. Humankind faces a most crucial mission; we must endeavour, on a global scale, to restore and improve our natural and social environments. In this environmental study, we will use context-dependent differential computation to analyse changes in various factors (temperatures, colours, level of CO_2, habitats, sea levels, coral areas, etc.). In this paper, we will discuss a global environmental computing methodology for analysing the diversity of nature and animals, using a large amount of information on global environments.

Keywords. Context-dependent differential computation, cross-cultural data, environmental data, data mining processes, context computing, environmental ICT, globalisation.

Introduction

To promote discussion on context-dependent differential computation, we organised a panel session on "Cross-cultural and Environmental Data Analysis in Data Mining Processes for a Global Resilient Society" during the 25th International Conference on Information Modelling and Knowledge Bases (EJC2015). The panellists were Professor Yasushi Kiyoki (panel moderator and chair), Professor Xing Chen, Professor Petchporn Chawakitchareon, Professor Virach Sornlertlamvanich, and Senior Researcher Anneli Heimbürger. Our paper is based on the presentations of the panellists' own viewpoints on the session topic.

Our paper is organised as follows. In Section 1, Professor Yasushi Kiyoki presents the Keio University SFC Global Environmental Systems Leaders Program (GESL), and a global environmental analysis based on a semantic associative computing system. In Section 2, Professor Xing Chen introduces high-dimensional data processing engines for cross-cultural and environmental data analysis and mining. In Section 3, Professor

[1]Corresponding author.

Petchporn Chawakitchareon presents a comparison of prediction methods for alum dosage use in water supply treatment processes. In Section 4, Professor Virach Sornlertlamvanich introduces the Hyper Local News Generation System. In Section 5, Senior Researcher Anneli Heimbürger introduces an analysis method for context-sensitive vocalisation among brown bears, based on Sensing-Processing-Actuating (SPA) architecture. Section 6 summarises our paper.

1. Global Environmental Analysis and Cross-cultural Multimedia Computing

Keio University SFC has started a collaborative program between itself and its international collaborating institutes, the Keio University Global Environmental System Leaders Program (GESL) (http://gesl.sfc.keio.ac.jp/en/) (Figure 1). The aim of this program is to make a genuine contribution to the international community, by developing a workforce of global environmental system leaders with the ability to discover solutions to multifaceted environmental issues, based on a firm foundation of science and technology, and clearly formulated social rules.

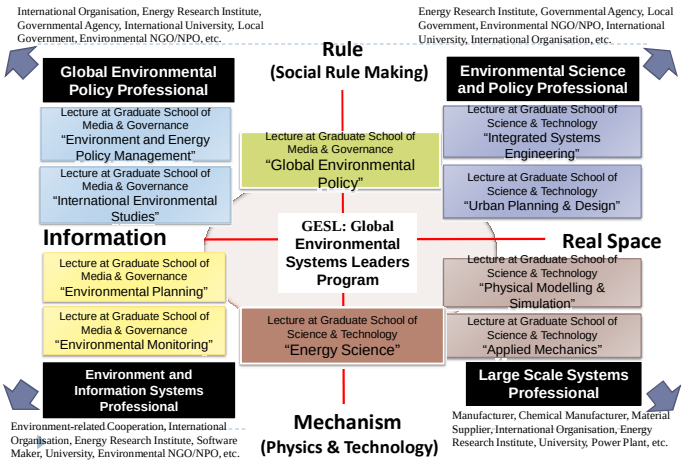

Figure 1. Keio University SFC Global Environmental Systems Leaders Program (GESL).

1.1. Environmental Analysis with the Semantic Associative Computing System

The important computation in environmental study is context-dependent-differential computation to analyse the changes of various situations (temperature, colour, CO_2, habitats, sea level, coral area, etc.). It is very important to memorise these situations and compute environmental changes in its various aspects and contexts, in order to investigate what is happening to the nature of our planet. We have almost infinite aspects and contexts regarding environmental changes, and a new method of analysis is needed to compute the differences in these situations. We propose a method of differential computing in our multi-dimensional world map [1, 2, 3]. We utilised a multi-dimensional computing model, the Mathematical Model of Meaning (MMM) [4, 5, 6], and a multi-dimensional space filtering method with adaptive axis adjustment

mechanism, in order to implement differential computing. Using this method, we are able to highlight important factors which change the natural environment. We also present a method of visualising the highlighted factors using our multi-dimensional world map.

We also propose a multimedia data mining system for global environmental analysis. In the design of such systems, one of the most important issues is how to search for and analyse media data (images, sound, movies, and documents), according to the user's contexts and environmental situations. We have introduced a semantic associative computing system based on our MMM. This system realises semantic associative computing in order to search for media data, and it is used to dynamically compute semantic correlations between keywords, images, sensing data, sound data, and documents in a context-dependent way. The main feature of this system is the realisation of semantic associative searching in the 2000-dimensional orthogonal semantic space, with semantic projection functions. This space was created for the dynamic computation of semantic equivalence or similarity between keywords and media data.

We have constructed a cross-cultural multimedia computing system for sharing and analysing different cultures with MMM functions, which are applied to cultural and multimedia data as a new platform of cross-cultural collaborative environments [1, 3]. This environment enables us to create a remote, interactive, and real-time cultural and academic research exchange among different countries. One of the most important applications of the semantic associative computing system is global environmental analysis, as shown in Figures 2 and 3, which aims to evaluate the various consequences of natural disasters in global environments. Our experiments' results have shown the feasibility and effectiveness of our semantic associative computing system, based on MMM, in global environmental analysis.

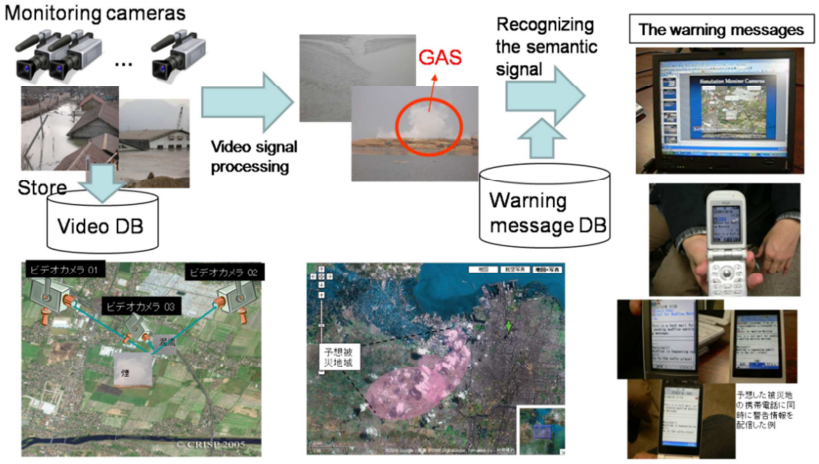

Figure 2. Environmental analysis with the semantic associative computing system.

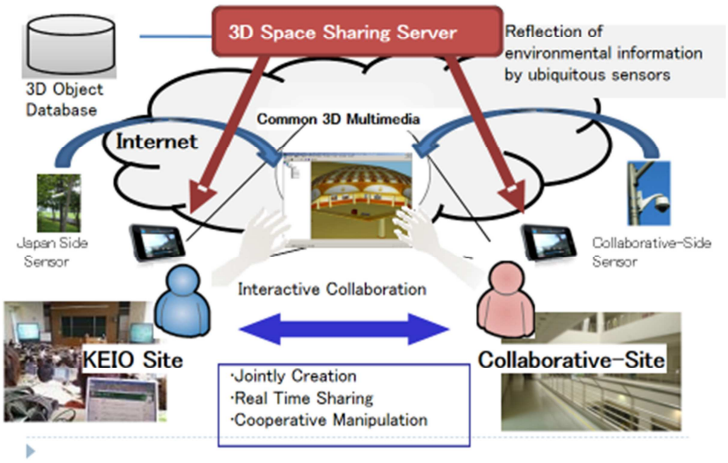

Figure 3. 3D cyberspace for environmental analysis and cross-cultural communication.

1.2. Spatio-Temporal and Semantic Computing

We have introduced the architecture of a multi-visualised and dynamic knowledge representation system, the 5D World Map System [1, 3], which is applied to cross-cultural multimedia computing. The basic space of this system consists of a temporal (1^{st}) dimension, spatial (2^{nd}, 3^{rd} and 4^{th}) dimensions, and semantic (5^{th}) dimension, representing a large-scale and multi-dimensional semantic space based on our semantic associative computing system (MMM). This space memorises and recalls various cross-cultural multimedia information resources with temporal, spatial, and semantic correlation computing functions, and realises a 5D world map for the dynamic creation of temporal, spatial, and semantic multiple views, which are applied to these resources.

We apply the dynamic evaluation and mapping functions of multiple views of temporal-spatial metrics, and integrate the results of semantic evaluation to analyse cross-cultural multimedia information resources. MMM is applied as a semantic associative search method for realizing the concept of "semantics" and “impressions” of cultural multimedia information resources, according to the "context". The main feature of this system is the creation of global maps, and views of cultural features, dynamically expressed in information resources (image, music, text, and video), according to the user’s viewpoints. Spatially, temporally, semantically, and impressionably evaluated and analysed cultural multimedia information resources are mapped onto a 5D time-series multi-geographical space. The basic concept of the 5D world map system is shown in Figures 4 and 5. The system, when applied to cross-cultural multimedia computing, visualises relations between different areas and times in terms of cultural aspects, by using dynamic mapping functions with temporal, spatial, semantic, and impression-based computations [2, 3, 7].

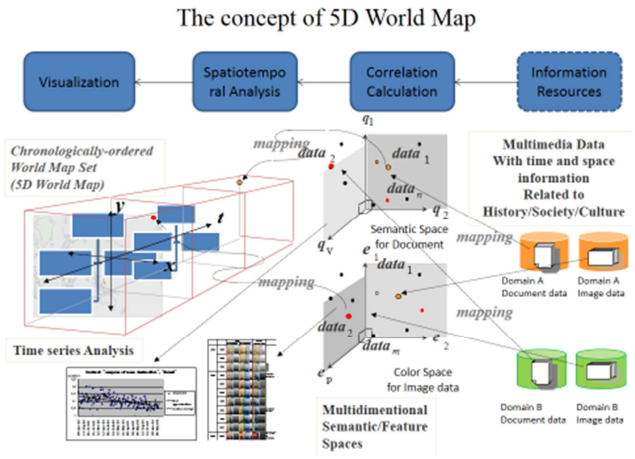

Figure 4. 5D world map system for worldwide viewing in global environmental analysis.

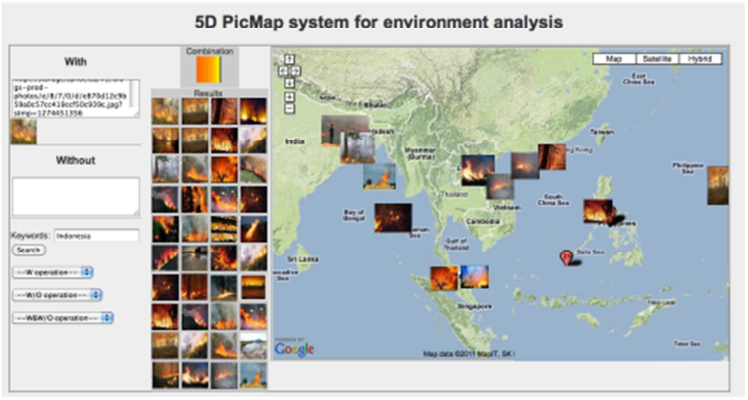

Figure 5. Global environmental analysis of forest fires in the 5D world map system.

In our future work, we will apply our multimedia computing system to new international and collaborative research and education for the realisation of mutual understanding and knowledge sharing of environmental and cross-cultural issues in the global view.

2. High-Dimensional Data Processing Engines for Cross-Cultural and Environmental Data Analysis and Mining

Widely used mobile devices which connect to the Internet, such as smartphones and wearable devices, are changing databases' data input methods from keyboard input, to mobile data transfer. At present, huge amounts of data are transferred from mobile devices to databases every day, and cloud technologies are also undergoing rapid development. Such progress, which alters the accumulation of data resources, leads to new environments of data analysis and mining; these come under the heading of "big-data" analysis and mining.

The essential purpose of such explorations is to find reasons, correlations, features, and so on in the data. Many clustering methods are developed by the division of data into different groups, which helps to identify features such as k-means, fuzzy c-means, quality threshold, kernel k-means, etc. [8]. Big-data structures are commonly the highest-dimensional data. This is one reason why data clustering algorithms cannot be standardised. That is, an algorithm may give the best result with one type of data set, but may fail, or deliver poor results, with other types, as illustrated in Figure 6.

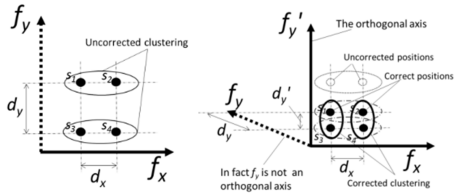

Figure 6. An example of clustering accuracy: in this case, the factors of data sets are not orthogonal to each other.

Figure 6 shows an image of clustering results. Two factors, f_x and f_y, are used for the clustering. Four data sets, s_1, s_2, s_3, and s_4, are mapped on a space constructed by f_x and f_y. The distances between the data sets d_x and d_y is used for the clustering. In the case that the distance value d_x is smaller the distance value d_y - that is, $d_x<d_y$, s_1, and s_2 are clustered in a group - s_3 and s_4 are clustered in the other group, as shown in Figure 6(a). However, it is common that factors extracted from data sets are not always orthogonal to each other. In this case, as shown in Figure 6(b), the factors f_x and f_y are not orthogonal to each other. Therefore, the values of d_y and d_y' are not equal to each other, and the value of d_y cannot be used for the clustering. The correct value of d_y' should be used on the space constructed by f_x and f_y'. As the distance value d_y' is smaller than the distance value d_x - that is, d_y' $<d_x$, the correct clustering, should cause s_1 and s_3 to be clustered in a group - s_2 and s_4 are clustered in the other group, as shown in Figure 6(b).

Research works are being carried out to create orthogonal space from data factors [9, 10]. A model referred to as the Mathematical Model of Meaning is presented in [4, 9]. In the model, an English dictionary is utilised, and an orthogonal space is created based on the appearance of the words used to define each entry in the dictionary. Another method has been proposed to create orthogonal spaces based on sample data sets [10]. This method is illustrated in Figure 7. An initial space is created, based on

extracted factors from data sets, as shown in Figure 7(a). Three steps are performed to create three orthogonal axes, as shown in Figure 7(b), (c), and (d).

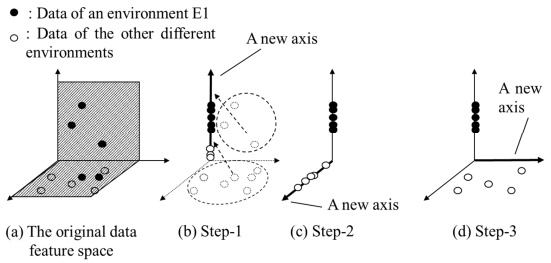

Figure 7. Automatic orthogonal space-creating method, based on data set samples.

In order to perform environmental data analysis, it is very important to compute changes in various aspects and contexts. As there are almost infinite aspects and contexts of environmental changes, a differential computing method is proposed [11] in order to realise a new analysing engine. This engine is used to compute differences in the discovery of actual aspects and contexts existing in the nature of our planet. The differential computing method is illustrated in Figure 8, where environmental data is represented as vectors $\mathbf{E}_i$ and $\mathbf{E}_j$. Differential vectors are generated during the computation, representing differences in features of data sets. At the same time, three vectors, $\mathbf{R}_c$, $\mathbf{R}_i$, and $\mathbf{R}_j$, are also generated, based on given threshold values. $\mathbf{R}_c$ is referred to as a *common vector*. $\mathbf{R}_i$ and $\mathbf{R}_j$ are referred to as *feature vectors*, with feature values that are greater than a given threshold, and smaller than a given threshold, respectively.

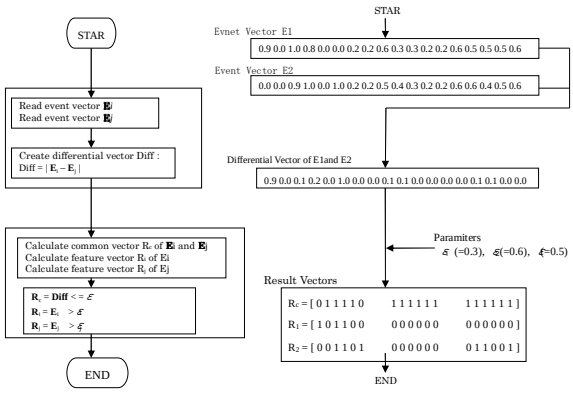

Figure 8. The differential computing method for multi-dimensional data sets.

The rapid progress of cloud, mobile devices and sensor technology has realized the large scale of global environmental data transfer and resource-accumulation in the world. “Global Environmental Analysis” is becoming one of the most important issues in global societies and communities connected in the world-wide scope. The innovative integration of large-scale multimedia environmental data management, and ubiquitous

computing, will lead to new methods of environmental analysis and cross-border communication. In our research work, we present the Global Environmental Analysis System. The basic idea of the system is illustrated in Figure 9. In the system, a global environmental-analysing engine, a global environmental event differential computing engine and a semantic associated engine are installed. Sensor data and cross-culture contents are stored in cloud databases. Following processing of differential calculations, orthogonal feature spaces are created for data analysis and mining.

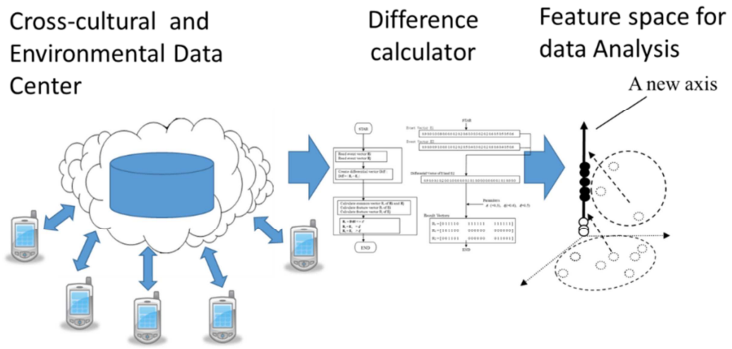

Figure 9. The Global Environmental Analysis System.

As a first step towards integration of the cross-cultural and environmental database with a differential computing engine and the data-analysing engine, we have developed a cloud application development platform referred to as FOCAPLAS [12]. The platform is composed of a formula parser, a formula calculator, a database management system, data storages, an input/output device, and interface specification interpretation equipment. As spreadsheet-based data modelling is supported, an application developer can build a cloud service by posting spreadsheet contents, just like posting document contents to a blog server. If developers wish to update their applications, they will be required to post new spreadsheet contents.

Our research is currently at an early stage in providing cloud services for cross-cultural and environmental data analysis and mining. Computer science areas, and environmental and global culture areas, are co-operating on an international basis. In our future work, we will provide cross-cultural and environmental data analysis and mining cloud services for a globally resilient society.

3. Alum Dosage Use in Water Supply Process Prediction by Decision Tree Forest Method and Genetic Programming

We will now present a comparison of prediction methods for alum dosage using in the water supply treatment process. A neural network is a common method, which has been used in many works. In this research, we compared the results from a decision tree forest to the results from neural networks. Seven input variables relating to the reaction of coagulation were used: turbidity, alkalinity, pH, conductivity, colour,

suspended solids, and NH_3-N. A new input variable was then generated by applying genetic programming. This new variable was used to improve the prediction result. The data for this research was collected from the Bangkhen Branch Office of the Metropolitan Waterworks Authority, Bangkok, Thailand, from 1 January to 31 December 2006. Our experiment's results showed that neural networks yielded more accuracy than the decision tree forest for seven input variables, but a decision tree forest with eight input variables yielded the highest accuracy compared to all other cases.

Coagulation is a process in which chemicals are added to water for the purpose of producing flocs from colloidal particles, and precipitating other contaminants. Coagulant dosage is non-linearly correlated to the removal of colloidal particles [13]. Hence, it is difficult to determine the optimal value for the dosage. The optimal coagulant dosage is an important parameter for effective control, monitoring and support of the process. Nowadays, most water supply plants determine coagulant dosage by using the Jar-Test method [14]. However, there are some disadvantages to this method; for example, it takes a long time to analyse samples in the laboratory (about three hours) [14], the process must be performed by skilled operators, and there is a limitation in feedback control. This is despite the fact that today there are a number of automatic coagulant control machines, such as the coagulometer [15], which passes a flat beam of light through water to determine the quantity of coagulant dosage which should be added through conversion of the quantity of received light, and streaming current detectors (SCDs) [16], which measure the negative charge of colloidal colour and turbidity particles in raw water. The appropriate amount of added coagulant dosage will adapt the charge of water to a neutralised state. Disadvantages of automatic coagulant control machines are the cost of equipment and operations, and a lack of adaptation to all levels of raw water quality. Therefore, in this paper, we propose a study to compare the various prediction methods of coagulant dosage, i.e. neural networks, support vector machine, single decision tree, and decision tree forest, in order to find the method which delivers the highest accuracy.

3.1. Methodology

The total data for our study comprised 2,014 records. To determine the most efficient method, we used a 10-fold cross-validation technique which divided the data into ten sets of size n/10 (n is number of records). Each set was tested by using the remaining sets as its training sets. Our comparing methods are described in detail as follows:

- Back-propagation neural networks: A neural network is a computation method which simulates signal transfers in the human brain. It has been used to predict coagulant dosage in many studies [13, 14, 17, 18, 19]. The structure of neural networks used in this paper consists of seven input nodes, one output node, and one hidden layer. The number of hidden nodes and iterations are tuned within the range of 5-20 and 10,000-50,000, respectively, to determine the best tuning for optimal performance.
- Support vector machine: This is an alternative learning method used for classification and regression. In this research, we chose two types of SVM model, namely the Epsilon-Support Vector Regression (ε-SVR), and the Nu-Support Vector Regression (ν-SVR). Four types of kernel function (linear function, polynomial function, radial basis function (RBF), and sigmoid function) were selected, in order to find the most accurate SVM.

- Single decision tree: Generally, a decision tree is used in classification problems. However, we can use the regression tree method to determine the coagulant dosage whose target value is continuous. In this case, we used a regression tree of which the predicted values are the mean value of target variables' falls in the leaf node. The value of the minimum number of nodes to be split, and the maximum tree levels, are tuned within the range of 2-50 and 10-500, respectively.
- Decision tree forest: This is an implementation of a random forest [20]. Used for regression problems, it is a collection of regression trees which are created by using bootstrap samples on the training data, and random feature selections in tree induction. Prediction is made by averaging the predictions of all trees in the forest. The number of trees in the forest, minimum number of nodes to be split, and maximum tree levels are tuned within the range of 10-500, 2-50, and 10-500 respectively.

3.2. Results

The results of the predictions using neural networks, SVM, single decision tree, and decision tree forest show that the decision tree forest gives the best result of prediction. A root mean square error of 2.71 can be achieved with the number of trees in the forest, minimum size node to split, and maximum tree levels, equal to 120, 3, and 40 respectively. The comparison of the root mean square error between methods is shown in Table 1. The best performance of neural networks can be achieved with 20 hidden nodes and 50,000 iterations. The root mean square error of this tuning is 4.84. The best performer for SVM is ε-SVR, using RBF. The values of parameter C, Gamma, and P are 800, 5, and 1, respectively. The root mean square error of this tuning is 3.66. The best performance of a single decision tree can be achieved with values of minimum size node to split, and maximum tree levels, of 8 and 20 respectively. The root mean square error of this tuning is 3.78.

Table 1. Comparison of RMSE (root mean square error)

	Neural Networks	**SVM**	**Single Decision Tree**	**Decision Tree Forest**
RMSE	4.09	3.66	3.78	2.37

Our results show that the decision tree forest method yielded the highest accuracy, compared to the other methods. The reason for the decision tree forest yielding a better result than the single decision tree is that the decision tree forest method employs multiple learners instead of a single learner, which can improve the accuracy in the same way as general ensemble methods.

Moreover, for the neural networks and support vector machine methods, the range of output varies highly. The results obtained from both methods are lower than the ones from the decision tree forest. In the future, we will utilise a new technique of using the decision tree forest method in order to extract the knowledge which can be used to predict the most suitable coagulant dosage in water supply plants.

4. Hyper Local NEWS Publishing: Collect, Analyse, and Visualise

The drastic advance in information technology is increasing the production, as well as the consumption, of data on an increasing number of web resources. In this work, we present an application for the delivery of local news to complement the deteriorating local newspaper, in order to promote data usage in rural areas. This can improve rural residents' livelihoods through the expression of their reputations on a hyper local news portal. News articles are collected from online public sources only; i.e. online news, government portals, Wikipedia, social media, collections etc., and specified by location. Articles on the same incident are automatically grouped using a 'keyword based text similarity' algorithm. The similarity is measured in two dimensions, content and location. Similar news from various publishers and sources is aggregated and classified into genres. Incidents are sorted based on reporting time, and can be automatically extended to a timeline of publishing. As a result, the news and information articles are automatically classified and presented by both location (province) and content (genre or news category).

In this information age, when people in cities need to access information, they jump on the Internet. Information overload is a major issue. It can take a long time to find relevant information among the millions of character strings on web pages. Sometimes we need to follow many links to find the information we need. However, people living in the countryside cannot access the Internet in the same way due to unreliable connections, and often cannot find news or information which is relevant to their communities.

But human information targeting of people living in the city is much different from the ones living in the countryside. People living in the countryside, a province outside the capital city, where the Internet is not such easily to get access as in the city, hardly search and look up information on the Internet neither. As a result, very often they face the situation of missing or insufficient information. Either the up to date news or necessary information related to their living and town are not delivered sufficiently to them.

We have automatically generated a hyper local news portal, which collects all information relevant to every province of Thailand, and classifies them according to provincial concerns and topics of interest. The idea is to serve people using a location-based approach. In each province, residents can find all the relevant information for their area on one page. People can easily find and learn about necessary or interesting things to do with their daily life.

4.1. Local News Collecting

The first step was to collect information from the Internet. We focused on four sources which provide local related information: Thai information web pages, knowledge web pages, social media, and news publishing web pages.

Information web pages provide factual information, and are mostly administered by educational institutions or government agencies [21]; for example, the website of the Mass Communication Organisation of Thailand, the website of the Department of Accelerated Rural Development, the website of the Office of National Buddhism, weather and stock market web pages, etc. [22]. Wikipedia is a source of collaborative knowledge. Social media data is collected from Facebook, Instagram, Google Plus, Twitter, and YouTube. Seven news web pages are also targeted. To collect the listed

data, we employ web crawling, Wikipedia Infobox harvesting, topic-based message extraction, and template-based crawling corresponding to the nature of the source of data. The data collected from these different procedures, which is ready to be used in the analysis process, comprises information, facts, social updates, and news. Figure 10 shows the data collection scenario.

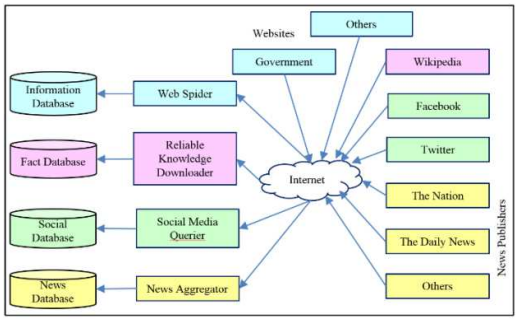

Figure 10. Data collection scenario.

Table 2 shows the size of data from the primary sources. The online news articles were collected during the period of January - September, 2014.

Table 2. Source portal for data collection with the size collected.

Category	Sources	Size
Online news		
Bangkok Post	www.bangkokpost.com	20,602 articles
Manager	www.manager.co.th	90,114 articles
Nation	www.nationmultimedia.com	43,275 articles
Thairath	www.thairath.co.th	40,420 articles
Thaipost	www.thaipost.net	26,699 articles
Daily news	www.dailynews.co.th	59,284 articles
Naewna	www.naewna.com	18,343 articles
Others		
Local product	www.cddopc.com/otopselect2013	3,207 products
Property	www.ddproperty.com	89,304 places
Job	www.jobthai.com	17,477 jobs
Government procurement	www.gprocurement.go.th	62,795 posts
Weather	www.openweathermap.org	5,937 days
Restaurants	www.wongnai.com	99,568 restaurants
Cinema	www.majorcineplex.com	1,091 movie schedules
Exchange rate	Exchange rate (THB<-> Other currency for 10 years)	599,433 exchange rate
Gold price	Gold price	3,702 days
Stock market index	SET index	1,095 days

4.2. Local News Analysis

After aggregating news and crawling all target information, NLP approaches are used to classify news categories, extract information, and analyse social media information. The classification process of news websites is shown in Figure 11. It consists of a word segmentation process, a named entity recognition process, and a news domain and

province classification process, together with term frequency ranking based on the Term Frequency-Inverse Document Frequency (TF/IDF) technique.

The name entity recogniser has been developed from an annotated corpus developed by [23] and [24]. A state-of-the art Thai morphological analyser, trained by ORCHID corpus [25] and TCL's lexicon [26], is used to obtain word boundaries and POS tags.

In this work, TF/IDF is used to identify keywords. We generate a Word Article Matrix (WAM) to create a map between an article, and a list of keywords from the article [27]. The WAM is generated by associating the value of the term frequency, in order to make the matrix ready to estimate the text similarity [28, 29]. Based on the named entity recognition, and TF/IDF-based WAM approaches, we extract the information related to the province and the category at an acceptably high accuracy. It is reported that the F-measure of the text classification approach falls between 85-100%, according to the genre [29].

From Wikipedia, the Infoboxes of any politicians and celebrities resident in a certain town are analysed according to the specific template.

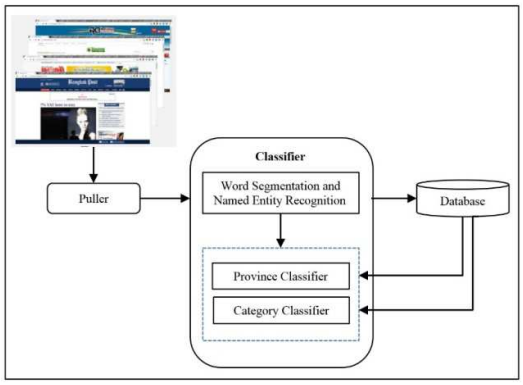

Figure 11. News classification methodology.

4.3. Local News Visualisation

Due to the amount of text to be classified, we manipulate it to be appropriately visualised in the portal. We apply document similarity-based ranking in order to organise the data into news categories such as educational news, art and performance news, political news, and so on. These categories are based on the major newspaper categories. Infoboxes are also used to visualise data in the portal. Text visualisation is implemented in order to present the information in the portal in a way that is clear, categorised, and easy to follow.

4.4. Hyper Local News Portal

The Hyper Local News Portal is finally available for public view. Information related to specific provinces is presented, so that users are able to scan, discover and learn about facts, trends, and hot or breaking news from any provinces or their own

hometown, on a single page. Figure 12 shows screenshots of the portal, where users can find categories of information related to provinces.

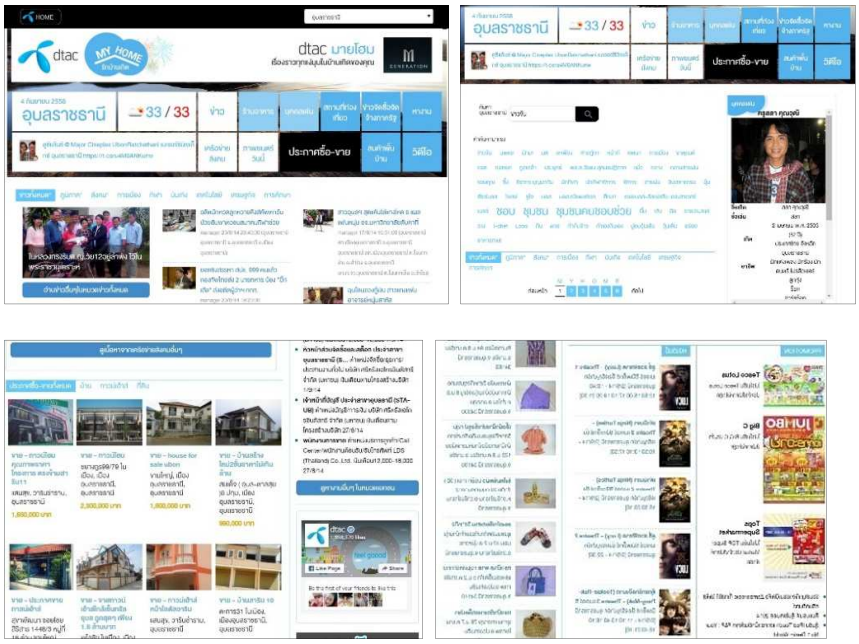

Figure 12. Screenshots of the Hyper Local News Portal.

The portal provides categorized information for all 77 Thai provinces. People in the province can search and discover all information related to their province in one portal. The portal of Hyper Local News can then serve the information for the local daily life i.e. news, restaurants, movies, shopping, weather, celebrities, and so on. News and information in the portal is as up-to-date and accurate as it is in the original pages.

5. SPA-architecture and Context-Sensitive Vocalisation among Brown Bears

Bears have captured our imagination for centuries. Ancient Finnish and Lappish myths and legends are probably one reason why these powerful animals are still held in great respect today in those countries, and referred to as the Kings of the Forests [30, 31]. Worldwide, according to our current knowledge based on DNA analysis, the bear taxonomy includes eight still living species: brown bears, polar bears, black bears, white-chested bears, sun bears, sloth bears, spectacled bears and pandas [32]. The focus of our study is on brown bears and their communication in certain situations and contexts.

Close monitoring, information collection, and analysis give us more precise information on bears' behaviour. Based on that information, we can create new knowledge on bears' biology and their environmental state. The behavioural and

communication schema of the bear seems to be very goal-oriented and situation-specific. If we can recognise context-dependent communication schemas, we will be able to create a lexicon of bear communication. This lexicon could be used by, for example, scientists, authorities, teachers, students, hikers, and, especially, citizens living in bear-rich areas.

It is widely believed that bears behave in an unpredictable way. However, as with all mammals, their behaviour is governed by a combination of genetic programming, and social and environmental factors.

The focus of our research is on context-based bear communication; specifically, bear vocalisation and body language in certain situations. We will introduce a context-based schema for brown bear communication research, which is based on the sensing, processing, and actuating (SPA) architecture (Figure 13) [33, 34, 35]. The system described here is at an early stage of implementation. The three main SPA phases are briefly described as follows.

Sensing: Finland's Ähtäri Zoo [36] offers unique opportunities to study contextual bear communication in a fixed, but quite expansive space. At the moment, there are four bears in the zoo. The movements and voices of individual bears are easily followed by means of GSP collars, web cameras, and human perception. Forests would, of course, provide us with an open research environment. However, forests better suits bear population studies carried out by the Natural Resources Institute, Finland [37]. Population study has long been the main contribution to bear research in Finland. Now, bear behaviour and communication studies are also being studied, for example at the University of Jyväskylä.

The main idea of our research is to identify groups of voice sequences which are typical for certain situations. The following context classes are used in the study: bear ID, which includes name, gender, and age; GPS position, which indicates the bear's location in the fenced area; and season, either spring, summer, autumn, or winter. Situations, as a context class, include: waking up from hibernation; cubs coming out from the den for the first time; a female bear teaching her cubs; cubs playing; friendly wrestling between cubs or adults; mating wrestles between two males; hunting and eating; defence of territory; and, finally, going into hibernation.

Processing: Voice signal classification consists of extracting and selecting physical and perceptual features from a voice signal. By using these features, it is possible to identify into which context class the voice is most likely to fit. Feature extraction is a process whereby a segment of a voice signal is characterised with a compact numerical representation. If the features extracted are carefully chosen, it is expected that they will contain relevant information from the input data. A desired task can then be performed using that reduced representation, instead of the full-sized input. Feature selection is the process of removing features from the set which are less important as regards the classification task to be performed. We study the feature extraction of context-dependent vocalisation of bears by means of the following features: temporal (features are calculated from the input waveform), spectral (features are computed from short-time Fourier transform of the input signal), perceptual (features are computed from the human perceptual model), and harmonic (features are computed from the sinusoidal harmonic model of the signal). Signal processing is carried out in the Matlab environment [38, 39].

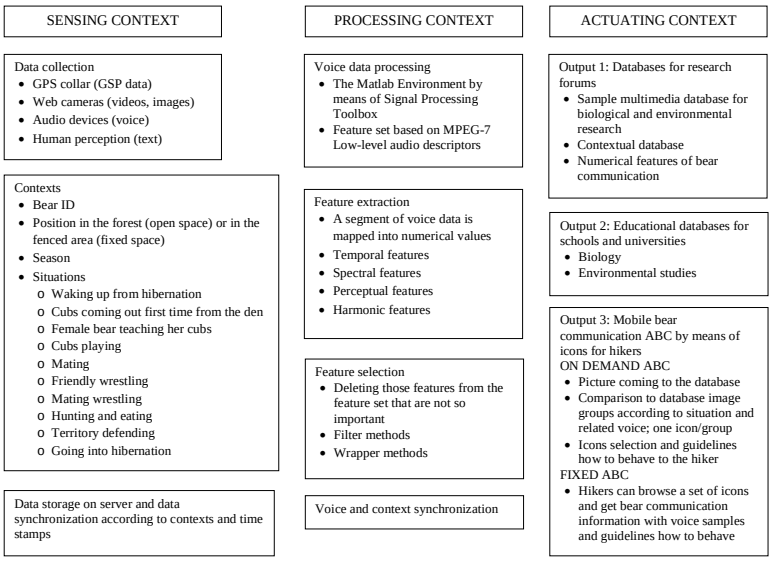

Figure 13. Context-based bear communication research schema.

Actuating: We describe here three examples of how the created multimedia bear database can be used, in addition to our bear vocalisation context analysis. The bear database includes voice and video sequences, still images and textual information, and numerical features of bear communication, together with contextual and temporal information. Firstly, the database provides valuable information for biological and environmental research forums, and for authorities. Secondly, it can also be used for biological and environmental education in schools and universities. Thirdly, by means of the database and icons, we can create a mobile bear communication lexicon for hikers. In “On Demand Lexicon”, the hiker can send a bear picture taken with her/his mobile phone to the server. The input picture is compared to database image groups, according to situations and related voices. An icon representing the most similar image group will be selected. This icon, with a voice sample, will be sent to the hiker, with some guidelines about how to behave. In “Fixed Lexicon”, hikers can browse a set of icons on their phone, and get bear communication information with voice samples and guidelines on how to behave.

6. Conclusions

In this paper, we have discussed a global environmental computing methodology for analysing the diversity of nature and animals, using a large amount of information on global environments. The important computation in environmental study is context-dependent-differential computing for analysing the changes of various situations in nature. In our paper, we have presented several points of view to global environmental computing methodology for analysing differences and diversity of nature and livings with a large amount of information resources.

References

[1] Y. Kiyoki, S. Sasaki, N. Nguyen Trang, N. Thi Ngoc Diep. *Cross-cultural Multimedia Computing with Impression-based Semantic Spaces, Conceptual Modelling and Its Theoretical Foundations*, Lecture Notes in Computer Science, Springer, pp. 316-328, March 2012.

[2] Y. Kiyoki. A *Kansei: Multimedia Computing System for Environmental Analysis and Cross-Cultural Communication*, 7th IEEE International Conference on Semantic Computing, keynote speech, Sept. 2013.

[3] S. Sasaki, Y. Takahashi and Y. Kiyoki. *The 4D World Map System with Semantic and Spatiotemporal Analyzers*, Information Modelling and Knowledge Bases, Vol.XXI, IOS Press, 18 pages, 2010.

[4] Y. Kiyoki, T. Kitagawa and T. Hayama. A metadatabase system for semantic image search by a mathematical model of meaning, *ACM SIGMOD Record*, vol. 23, no. 4, pp. 34-41, 199

[5] Y. Kiyoki, T. Kitagawa and T. Hayama. *A Metadatabase system for semantic image search by a mathematical model of meaning*, Multimedia Data Management -- using metadata to integrate and apply digital media, McGrawHill (book), A. Sheth and W. Klas (editors), Chapter 7, 1998.

[6] Y. Kiyoki and S. Ishihara. *A Semantic Search Space Integration Method for Meta-level Knowledge Acquisition from Heterogeneous Databases*, Information Modeling and Knowledge Bases (IOS Press), Vol. 14, pp. 86-103, May 2002.

[7] T. Suhardijanto, Y. Kiyoki, and A. Ridho-Barakbah. *A Term-based Cross-Cultural Computing System for Cultural Semantics Analysis with Phonological-Semantic Vector Spaces*, Information Modelling and Knowledge Bases XXIII, pp. 20-38, IOS Press, 2012.

[8] R. Xu and D. Wunsch. *Survey of clustering algorithms*, IEEE Transactions on Neural Networks, Vol. 16, No. 3, pp. 645-678, 2005.

[9] X. Chen and Y. Kiyoki. *A dynamic retrieval space creation method for semantic information retrieval*, Information Modelling and Knowledge Base (IOS Press), Vol. XVI, pp. 46-63, 2005.

[10] Y. Kiyoki and T. Kitagawa. *A semantic associative search method for knowledge acquisition*," Information Modelling and Knowledge Base (IOS Press), Vol. VI, pp. 121-130, 1995.

[11] Y. Kiyoki and X. Chen. *Contextual and Differential Computing for the Multi-dimensional World Map with Context-specific Spatial-temporal and Semantic Axes*, Information Modelling and Knowledge Base (IOS Press), Vol. XXV, pp. 82-97, 2014.

[12] X. Chen and K. Shiohara. *FOCAPLAS – A platform for cloud application development and running support*, Information Modelling and Knowledge Base (IOS Press), Vol. XXVI, pp. 61-76, 2014.

[13] N. Valentin, T. Denoeux and F. Fotoohi, *Modelling of Coagulant Dosage in a Water Treatment Plant*, https://www.hds.utc.fr/~tdenoeux/dokuwiki/_media/en/congres/eann99.pdf, 1999.

[14] M. Barker, M. Nickels and H. Mayfield. Optimizing Drinking Water Treatment Using Neural Networks. http://www.nku.edu/~norsci/issue1/2003-1barker, 2003.

[15] Degremont. *Water Treatment Handbook*. Fifth Edition. Wiley, New York; N.Y., http://www.degremont.com/en/about-us/publications/handbook/water-treatment-handbook/, 1979.

[16] F. Bernazeau, P. Pierrone and J.P. Duguet. Interest in Using a Streamline Current Detector for Automatic Coagulant Dose Control. *Water Supply*. 87-96. 1992.

[17] E.S. Nahm, S.B. Lee, K.B. Woo, B.K. Lee and S.K. Shin. *Development of an Optimum Control Software Package for Coagulant Dosing Process in Water Purification System*. SICE '96, Proceedings of the 35th SICE Annual Conference, International Session Papers, Tottori University, July 24-26, 1157-1161, 1996.

[18] T.H. Han, E.S. Nahm, K.B. Woo, C.J. Kim and J.W. Ryu. *Optimization of Coagulant Dosing Process In Water Purification System*, SICE '97, Proceedings of the 36th SICE Annual Conference. International Session Papers, Tokushima, July 29-31,1105-1109, 1997.

[19] M.G. Chun, K.C. Kwak and J.W. Ryu. *Application of ANFIS for Coagulant Dosing Process in a Water Purification Plant*. IEEE International Fuzzy Systems Conference Proceedings, Seoul, Korea, August 22-25, 1999.

[20] L. Breiman. *Random Forests*. Machine Learning, 45, 5-32, 2001.

[21] Types of web sites-a categorization based on content [Online]. Available: http://www.webdevelopersnotes.com/.

[22] Thai Government websites. Available: http://www.cabinet.thaigov.go.th/webguide.htm.

[23] T. Theeramunkong, M. Boriboon, C. Haruechaiyasak, N. Kittiphattanabawon, K. Kosawat, C. Onsuwan, I. Siriwat, T. Suwanapong, and N. Tongtep. Thai-nest: A framework for Thai named entity tagging specification and tools, Proceedings of International Conference On Corpus Linguistics (CILC), 2010.

[24] C. Kruengkrai, V. Sornlertlamvanich, W. Buranasing, and T. Charoenporn. *Semantic Relation Extraction from a Cultural Database*, Proceedings of Workshop on South and Southeast Asian NLP, Proceedings of International Conference on Computational Linguistics (COLING), 2012.

[25] V. Sornlertlamvanich, T. Charoenporn, and H. Isahara. ORCHID: *Thai Part-Of-Speech Tagged Corpus*, Technical Report TR-NECTEC-1997-001, NECTEC (1997).
[26] T. Charoenporn, C. Kruengkrai, V. Sornlertlamvanich, and H. Isahara. *Acquiring semantic information in the TCL's computational lexicon*, Proceedings of the Fourth Workshop on Asia Language Resources, 2004.
[27] V. Sornlertlamvanich, E. Pacharawongsakda, and T. Charoenporn. *Understanding Social Movement by Tracking the Keyword in Social Media*, Proceedings of Multiple Approaches Lexicon (MAPLEX), 2015.
[28] T. Murakami, Z. Hu, S. Nishioka, A. Takano, and M. Takeichi. *An Algebraic Interface for GETA Search Engine*, Proceedings of Program and Programming Language Workshop, 2004.
[29] P. Jotikabukkana, V. Sornlertlamvanich, O. Manabu, and C. Haruechaiyasak. *Effectiveness of Social Media Text Classification by Utilizing the Online News Category*, Proceedings of International Conference on Advanced Information: Concepts, Theory and Application (ICAICTA), 2015.
[30] J. Pentikäinen. *On Bear'sHeels*, Helsinki University Press, Helsinki, 2005.
[31] R.E. Bieder. *Bear*, Reaktion Books Ltd, London, UK, 2005.
[32] F. Hailer et al. Nuclear Genomic Sequences Reveal that Polar Bears are on Old and Distinct Bear Lineage, *Science* **336** (2012), 344-347.
[33] A. Heimbürger and S. Kärkkäinen. *On Modelling Context-based Vocalization among Brown Bears*. In: IBA 2014, 23th International Conference on Bear Research and Management, Thessaloniki, Greece, 5-11 October 2014, Book of Abstracts, p. 123. ISBN: 978-960-7742-49-0
[34] A. Heimbürger. *On Modelling Context-Sensitive Communication based on SPA-Architecture. Case study: Bear Vocalization*. In: T. Tokuda, Y. Kiyoki, H. Jaakkola and N. Yoshida (eds.) Frontiers in Artificial Intelligence and Applications, Vol. 260 Information Modelling and Knowledge Bases XXV. Amsterdam: IOS Press. Pp. 255-263, 2014.
[35] A. Heimbürger. *On Modelling Context-Sensitive Communication based on SPA-Architecture*. In: Y. Kiyoki and T. Tokuda (eds.) Proceedings of the 23rd European-Japanese Conference on Information Modelling and Knowledge Bases (EJC2013), Nara, Japan, June 3-7, 2013.
[36] *Ähtäri Zoo*, <URL=http://www.ahtarinelainpuisto.fi/in_english>.
[37] *Natural Resources Institute* Finland, <URL= http://www.luke.fi/en>.
[38] *MATLAB*, <URL= http://en.wikipedia.org/wiki/MATLAB>.
[39] *Signal Processing Toolkit*, <URL= http://www.mathworks.se/products/signal/>.