3. Prediction, Forecasting, and Chance Discovery

Yutaka Matsuo

Cyber Assist Research Center, National Institute of Advanced Industrial Science and Technology, Aomi 2-41-6, Tokyo 135-0064, Japan
email: y.matsuo@aist.go.jp

Summary.

This chapter addresses the relation and difference between prediction, forecasting, and chance discovery. Prediction and forecasting have a long history. So far, many studies have been devoted to prediction and forecasting. However, in complex real-world systems, contrary to scientific laws, it is sometimes very difficult to predict the future. In such situations, model creation, model selection, and parameter fitting are all important in the complex changing real world. Chance discovery targets three aspects that prediction and forecasting methods have not shed light on, i.e. emphasis on model and variable creation and discovery, emphasis on rare events, and emphasis on human and computer interaction.

3.1 Introduction

This chapter addresses the relation and difference between prediction, forecasting, and chance discovery. Prediction and forecasting have a long history. From remote history, such as in ancient Greece, man demonstrated the desire to predict the future and understand the past; these desires motivated the search for laws that explain behavior of observed phenomena.

Scientific discoveries are sometimes verified through prediction: prediction of the planet Neptune's existence by Leverrier, prediction of deviation of light by Einstein, prediction of the helical structure of DNA by Watson and Crick, etc. [3.23]. Prediction has a very strong force of argument. So far, many studies have been devoted to prediction and forecasting. However, in complex real-world systems, contrary to scientific laws, it is sometimes very difficult to predict the future. The difficulty of prediction depends on the degree of freedom and complexity of the system; if too many parameters should be fixed, it is impossible to make a precise prediction. If the evolution law amplifies initial uncertainty too rapidly, one can not make long-term predictions.

In such situations, choice of a prediction model strongly affects the prediction performance. A model which works well in one case might not work well in other cases. Therefore, model creation, model selection, and parameter fitting are all important in the complex changing real world.

In contrast to the long history of prediction and forecasting, chance discovery is a brand-new research field; formally it began in 2000 (although many essential pieces of research had already begun in the late 1990s). Chance discovery targets aspects that prediction and forecasting methods have not shed light on: rather, those

aspects that prediction and forecasting had considered as given. In this section, differences are classified into three categories: emphasis on model and variable creation and discovery, emphasis on rare events , and emphasis on human and computer interaction. Conventional prediction and forecasting methods presume that a user (of the method) knows already which variables to predict, and which variables should be cast into the methods. For example, an investor wants to know the trend of a certain stock price based on the history of the price or data of other stock prices and economic indices; a marketer wants to predict sales of a product based on the sales history; a traveler wants to know tomorrow's weather based on the history of weather changes and current weather. However, sometimes one does not know which variable to predict: one can imagine a woman who is not aware of the risk of great earthquakes living in a quake-prone area, or a man who is not aware of the potential chance of developing a new hit product. These people do not know which variable to predict. In the real world, often in very important situations, we are not aware of which variables to predict, and which variables to cast into prediction methods.

Furthermore, ordinal prediction and forecasting methods postulate the existence of a coherent model behind data. If we assume coherence, many prediction and forecasting methods work very well. Certainly, scientific laws are very coherent. However, in the real world, sometimes it is not reasonable to assume coherence. Social and economic relationships are constantly changing. New products appear day by day. The Internet emerged globally, completely changing our way of life and business activities. Greenhouse gases have become a problem on a world-wide scale, resulting in the regulation of greenhouse-effect gas emissions, and leading to a new market for ecological hybrid cars. In such a real world, the assumption of a coherent model sometimes does not hold. Rather, we should develop methodology in the structurally changing world in which we live.

The following section makes a brief survey of prediction and forecasting methods. Knowing that prediction and forecasting constitutes a long-studied area, we cover only limited aspects of that field. Further information can be found, for example, in [3.28, 3.15, 3.7, 3.5]. Recent advances in data-mining methods open a new direction to prediction and forecasting. After overviews presented here, we will discuss the difference and relevance between prediction/forecasting and chance discovery in Sect.3.3. Section 3.4 is devoted to one model which we think captures the changing world: the small world. Some surveys and discussions are made there with regard to the small world .

3.2 Existing Method of Prediction and Forecasting

3.2.1 Time-Series Prediction

Weigend and Gershenfeld indicate that time-series analysis has three goals: forecasting, modeling, and characterization [3.28]. Forecasting is also called predicting; it aims at accurately predicting the short-term evolution of a system. (Prediction is

also referred to as estimating unobservables, for example of an RNA structure or of a VLSI circuit.) The goal of modeling is to find a description that accurately captures features of the system's long-term behavior. The third goal, system characterization, attempts with little or no a priori knowledge to determine fundamental properties, such as the number of degrees of freedom of a system or the amount of randomness.

Before the 1920s, forecasting was done by simply extrapolating the series through a global fit in the time domain. The beginning of 'modern' time-series prediction might be set at 1927 when Yule invented the autoregressive technique in order to predict the annual number of sunspots. His model predicted the next value as a weighted sum of previous observations of the series [3.31].

According to [3.28], two crucial developments occurred around 1980 due to general availability of powerful computers. The first development was state-space reconstruction by time-delay embedding. The second development was emergence of the field of machine learning; it was able to adaptively explore a large space of potential models. With the shift in artificial intelligence from rule-based methods toward data-driven methods, the field was ready to apply itself to time-series.

3.2.2 ARMA Model

Linear time-series models are one of the most simple predictive models; they can be understood in great detail and are straightforward to implement. ARMA models have dominated all areas of time-series analysis and discrete-time signal processing for more than half a century [3.28]. Two crucial assumptions will be made: the system is assumed to be both linear and stationary.

Assume that we are given an external input series $\{e_t\}$ and seek to modify it to produce another series $\{x_t\}$. In the MA (moving average) model, the present value of x is influenced by the present and N past values of the input series e:

$$x_t = \sum_{n=0}^{N} b_n e_{t-n} = b_0 e_t + b_1 e_{t-1} + \ldots + b_N e_{t-N}.$$

In the AR (autoregressive) model, some feedback is considered:

$$x_t = \sum_{m=1}^{M} a_m x_{t-m} + e_t.$$

Depending on the application, e_t can represent either a controlled input to the system or noise.

The ARMA model is a combination of the AR and MA models; the ARMA(M, N) model is stated as

$$x_t = \sum_{m=1}^{M} a_m x_{t-m} + \sum_{n=0}^{N} b_n e_{t-n}.$$

We can estimate coefficients of the AR(M) model from the observed correlational structure of a signal. Estimation of the coefficients can be viewed as a regression problem: expressing the next value as a function of M previous values. This

can be done by minimizing squared errors: the parameters are determined such that the squared difference between the model output and the observed value, summed over all time steps in the fitting region, is as small as possible. Standard techniques exist, often expressed as efficient recursive procedures, for finding MA and ARMA coeficients from observed data.

Historically, an important step beyond linear models for prediction was taken 20 years ago; it used two linear functions instead of one globally linear function. This threshold autoregressive model (TAR) is globally non-linear. Such non-linear models significantly expand the scope of possible functional relationships for modeling time series, but this benefit comes at the expense of simplicity. One solution to this is in a connectionist framework.

3.2.3 Pattern Recognition

One new developing method of forecasting is through pattern imitation and recognition [3.22]. Consider the time series as a vector

$$\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\},$$

where n is the total number of points in the series. The current state is represented as y_n. One possible simple method of prediction is based on identifying the closest neighbor of y_n in the past data, say y_j, and predicting y_{n+1} on the basis of y_{j+1}. This simple approach may be extended by taking an average prediction based on a set of nearest neighbors. The definition of the current state of a time series may be extended to include more than one value. Optimal state size must be determined experimentally on the basis of achieving minimal errors on standard measures.

Consider again the time series $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$. A segment in the series may be defined as a difference vector $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_{n-1})$, where $\sigma_i = y_{i-1} - y_i \ (\forall i,\ 1 < i < n-1)$. A pattern contains one or more segments and may be visualized as a string of segments

$$\boldsymbol{\sigma} = (\sigma_i, \sigma_{i+1}, \ldots, \sigma_h)$$

for given values of i and h, where $1 < i < h < n-1$. If we choose to represent the pattern more simply, we encode the time series $\boldsymbol{y}$ as a vector of change in direction: a value y_i is encoded as 0 if $y_i - 1 < y_i$, as a 1 if $y_i - 1 > y_i$, and as a 2 if $y_i - 1 = y_i$. A pattern in the time series may now be represented as

$$\boldsymbol{\rho} = (b_i, b_{i-1}, \ldots, b_h).$$

In this approach, time-series forecasting refers to the process of matching a current state of the time series with its past state. Success in correctly predicting the series depends directly on the pattern-matching algorithm. Also, the size k has an important impact on error minimization and correct prediction. The match itself is sometimes not exact and can be done by a fuzzy matching algorithm.

Similarly, aside from fuzzy methods, a large number of studies have been done for forecasting using neural networks, genetic algorithms, and Markov models.

3.2.4 Information Between the Past and the Future

In [3.4], Bialek et al. say that the only components of incoming data that present the possibility of being useful are those that are predictive. It makes sense to isolate the predictive information from non-predictive information. Learning a model to describe a data set can be seen as an encoding of that data; the quality of this encoding can be measured using information-theory concepts.

From the information-theory perspective, past data T provides information about future data T'. We can write the average of this predictive information as

$$I_{\mathrm{pred}}(T,T') \leq \left\langle \log_2 \frac{P(x_{\mathrm{future}}|x_{\mathrm{past}})}{P(x_{\mathrm{future}})} \right\rangle \tag{3.1}$$

$$= S(T) + S(T') - S(T+T'), \tag{3.2}$$

where $\langle \cdots \rangle$ denotes an average over the distribution; $S(T) = -\langle \log P(x_{\mathrm{past}}) \rangle$ is the entropy of observations on a window of duration T. From the formula above, we can view $I_{\mathrm{pred}}(T,T')$ as either the information that a data segment of duration T provides about the future of length T', or the information that a data segment of duration T' provides about the immediate past of duration T.

If we have been observing a time series for a long duration T, then the total amount of data we have collected is measured by the entropy $S(T)$. Under some assumptions, we can write $S(T) = S_0 T + S_1(T)$; of the total information we have taken in by observing x_{past}, only a vanishing fraction is relevant to the prediction:

$$\lim_{t\to\infty} \frac{\text{Predictive information}}{\text{Total information}} = \frac{I_{\mathrm{pred}}(T)}{S(T)} \to 0.$$

In this sense, most of what we observe is irrelevant to the problem of predicting the future.

3.2.5 Data-Mining Methods

Time-series data has been recently studied in the context of data mining. Many methods attempt to find frequent patterns in time-series data (e.g. [3.10]). APRIORI is one of the most well-known methods to find association rules

$$X \to Y.$$

Agrawal and Srikant introduced the sequential pattern-mining problem in [3.24]. Many methods which are based on the APRIORI property [3.1] have been proposed for mining sequential patterns (e.g. [3.2, 3.24, 3.9]).

Han *et al.* studied periodicity search, that is, search for cyclicity in time-related databases [3.12]. They found segment-wise periodicity in the sense that only some of the segments in a time sequence have cyclic behavior. For example, Laura may read a newspaper at 7:00 to 7:30 every weekday morning, but may do all sorts of things afterwards.

In contrast to mining frequent patterns or periodical patterns, several studies focus on rare events. Weiss proposed a method to predict extremely rare events such

as hardware-component failures in the AT&T network [3.29, 3.30]. Their system, called Timewaver, is a genetic-based machine learning system for predicting events. Following their description of the event-prediction problem, a prediction occurring at time t, Pt, is said to be correct if a target event occurs within its prediction period. The system searches the solution space using a genetic algorithm. Prediction rules are encoded into each individual. The rule is for example: if two (or more) A events and three (or more) B events occur within an hour, then predict the target event. They use precision and recall to evaluate a solution. Recall is the percentage of target events correctly predicted and precision is the percentage of times that a target event is predicted and actually occurs. The evaluation function is based on both precision and recall. The F-measure, which is used in information retrieval, is used as the evaluation function:

$$f = \frac{(\beta^2 + 1)\text{precision} \times \text{recall}}{\beta^2\text{precision} + \text{recall}}.$$

Instead of usual direct association, Tan et al. introduced the concept of indirect association between items [3.26]. They believed that some of the infrequent item sets may provide useful insight about the data. Consider a pair of items, (a, b), that seldom co-occur together in the same transaction. If both items are highly dependent on the presence of another item set, Y, then the pair (a, b) is said to be indirectly associated via Y. In market basket data, this method can be used to perform competitive analysis of products. For text documents, indirect association between a pair of words often corresponds to synonyms, antonyms, or words that are present in the different contexts of another word. This method is also used for mining Web-usage data [3.25].

Domeniconi et al. attempted prediction of significant events from sequences of data with categorical features [3.6]. Co-occurrence analyses of events are done by means of singular value decomposition of examples constructed from data. Starting with an initial rich set of features, they clustered features based on correlation. The resulting classifier was expressed in terms of a reduced number of examples; thereby, predictions can be performed efficiently.

In [3.23], catastrophic events are discussed such as the rupture of composite materials, great earthquakes, turbulence, abrupt changes of weather regimes, financial crashes, and human parturition. A central property of such complex systems is the possible occurrence of coherent large-scale collective behaviors with a very rich structure, resulting from repeated non-linear interactions among their constituents. These systems in natural and social sciences exhibit rare and sudden transitions, which occur over time intervals that are short compared to the characteristic time scales of their posterior evolution. Such extreme events express, more than anything else, underlying forces. In case of the rupture of materials, the fracture process depends strongly on the degree of material heterogeneity: if the disorder is too small, then the precursory signals are essentially absent and prediction is impossible. If heterogeneity is large, rupture is more continuous.

Finally, Last et al. [3.14] introduced new aspects and difficulties of time-series databases (TSDB). The process of knowledge discovery in TSDB includes cleaning

and filtering of time-series data, identifying the most important predicting attributes, and extracting a set of association rules that can be used to predict future time-series behavior. They used a fuzzy approach to express extracted rules in natural language.

3.3 Difference between Prediction/Forecasting and Chance Discovery

As seen above, myriad frameworks have been developed for predicting the future, including statistical, pattern-recognition, and data-mining algorithms. The major concern of chance discovery is also in the future, e.g. predicting earthquake occurrence, developing new merchandise, and planning new strategies. However, chance discovery targets those aspects that prediction and forecasting methods have not shed light on. Rather, on what in prediction/forecasting had been considered as given. We will discuss three aspects of chance discovery: model and variable creation and discovery, rare events, and human–computer interaction.

3.3.1 Emphasis on Model/Variable Creation and Discovery

Conventional prediction/forecasting methods postulate the existence of a coherent model behind the data. If we assume the coherence, many prediction/forecasting methods work very well. Certainly, scientific laws are very coherent. However, in the real world, is it reasonable to assume coherence? In the real world, the assumption of a coherent model often does not hold. Rather, we should develop methodology for a structurally changing world that resembles the one in which we live.

In real life, such as in the business world, human networks, social development, and so on, it happens very often that the structure of the system changes at some points. In [3.11] the way in which a little thing can cause a big structural change is discussed. Not only does a system evolve gradually as time passes, but the system may also completely change its structure at some points. This is due to large-scale collective behaviors with a very rich structure and repeated non-linear interactions among its constituents. In such a situation, conventional prediction and forecasting methods are not as effective as in a stable situation. In fact, when we face dramatic structural change, we may not be able to predict the future. It is very important to grab what happens, and find which variables to focus on.

Therefore, chance discovery is not concerned so much with predicting the precise values of some variables in the future. Although such prediction is very important in a stable situation, it is not effective in dynamic situations. Knowing what is happening, determining which variables to monitor, and creating a new model are of great importance.

Model selection is a key issue in prediction/forecasting. There are some heuristics to find the proper model, such as Akaike information criteria (AIC) or minimal description length criteria. In the context of data mining, feature selection is also an important process, which selects informative attributes. However, what we mention here includes a big change of the model based on complex dynamics.

3.3.2 Emphasis on Rare Events

Rare events sometimes have a very large impact on social, economic, and business worlds. It is relatively easy to obtain knowledge about a frequent pattern, and thus it can be understood well. In this sense, a frequent pattern does not have large information if we assume that the a priori probability is modeled by common awareness of the event: if all competitors of a company know about an event, the information can not be a powerful strategic card.

On the other hand, rare events are not easy to recognize and use for decision making. Events with low frequency are sometimes neglected; thus they have much information. If most competitors of a company do not know about an event, it can present opportunities for the company.

Ordinary statistical methods are very useful if a model is assumed and the number of samples is large. However, these methods are not proper for rare events. (Note that there are some techniques to analyze rare events statistically [3.8].) If the number of samples is small, it is generally not statistically supported. Chance discovery focuses on the tail of the distribution. Even if a large number of samples are collected, the tail exists and sometimes the tail is a good source of information.

The above discussion is based on information in Shannon's sense [3.21]. That is, we discard the meaning of the event and only focus on the probability of an event. However, in the real world, we must also consider the impact of the event. Prediction of a big earthquake with low probability is important, but prediction of an event with low probability and a low impact has no merit. Therefore, whether an event has an impact or not is an essential aspect.

When we deal with rare events, it is not practical to consider the meaning and impact of every rare event beforehand because such rare events can emerge in a variety of ways. It is essential to use computer calculation to reduce the number of rare events which *might be* important.

3.3.3 Emphasis on Human and Computer Interaction

The third point of difference is that chance discovery is thus exploiting the future with the aid of humans.

Some rare events are simply noise, while others indicate great impact. It is completely impossible to fully automate the judgement of rare events. To understand the rare event, it is necessary to have a large amount of background knowledge. To implement a computer with a large amount of background knowledge is virtually impossible, as much artificial intelligence research has shown. Therefore, human and computer interaction is essential, which is discussed below.

In prediction and forecasting methods, it is assumed that a user (of the method) knows already which variables are to be predicted, and which variables are to be used (including the case where a part of a large number of variables are used). For example, an investor wants to know the trend of a certain stock price; a marketer wants to know how the sales will be; a traveler wants to know tomorrow's weather. However, how about those who are not aware of the risk of great earthquakes living

in a quake-prone area? How about those who are not aware of the potential chance of developing a new hit product? In the real world (and often in very important situations), we are not aware of which variables to predict and which variables should be used.

Therefore, it is important to suggest new variables to humans. Textual information is a good source of information to provide humans with new aspects of targeting data because natural language has an extremely large number of dimensions. It is very often the case that humans can discover a new variable to predict through the stimuli of language. In addition, visualization and communication are both very important aspects for aiding humans' creativity, which is described in detail in Chap. 6.

3.3.4 Relevance of Prediction/Forecasting and Chance Discovery

Although prediction and forecasting and chance discovery have different aspects based on different presuppositions, they are not exclusive. Rather, they are complementary. To predict the future, it is very important to understand the events; sometimes we must invent the model and variables. Chance discovery focuses on the process of understanding data and model-creation. Model selection, parameter fitting, and hypothesis verification follow this understanding and model-creation stage.

Actually, commonly used methods are the combination of *KeyGraph* and a statistical hypothesis test: *KeyGraph* is first used to understand the data and to create a hypothesis. Then, statistical prediction methods are used to evaluate the hypothesis.

3.4 Importance of Structural Information for Rare Events

Though it is very difficult to predict the future with structural change, some recent research shows promising results. One method for addressing structural changes is to concentrate on the network structure of data, and discover which node might cause a great structural change. *KeyGraph* is an algorithm to visualize the data and provide an insight to the future, especially on the rare events if they co-occur with multiple frequent clusters. The details of *KeyGraph* are given in Chap. 18.

The same idea can be grasped in other structural analysis: small worlds. Strength of weak ties is known in social psychological science. Centrality in a network is another example of measuring what is important and what is not.

3.4.1 Small Worlds

Graphs that occur in many biological, social, and man-made systems are often neither completely regular nor completely random, but have instead a 'small world' topology in which nodes are highly clustered yet the path length between them is small [3.27]. For instance, if one is introduced to someone at a party in a small world, one can usually find a short chain of mutual acquaintances that connect. In

the 1960s, Stanley Milgram's pioneering work on the small-world problem showed that two randomly chosen individuals in the USA are linked by a chain of six or fewer first-name acquaintances (in the scope of their experiments), known as 'six degrees of separation'[3.19]. Watts have shown that a social graph (a collaboration graph of actors in feature films), a biological graph (a neural network of the nematode worm C. Elegans), and a man-made graph (the electrical power grid of the western USA) all have a small-world topology [3.27]. The World Wide Web also forms a small-world network [3.3].

To formalize the notion of a small world, Watts define the clustering coefficient and the characteristic path length [3.27]:

- The characteristic path length, L, is the path length averaged over all pairs of nodes. The path length $d(i, j)$ is the number of edges in the shortest path between nodes i and j.
- The clustering coefficient, C, is a measure of the cliqueness of the local neighborhoods. For a node with k neighbors, then at most ${}_kC_2 = k(k-1)/2$ edges can exist between them. The clustering of a node is the fraction of these allowable edges that occurs. The clustering coefficient, C, is the average clustering over all nodes in the graph.

Watts define a small-world graph as one in which $L \geq L_{\rm rand}$ (or $L \sim L_{\rm rand}$) and $C \gg C_{\rm rand}$, where $L_{\rm rand}$ and $C_{\rm rand}$ are the characteristic path length and clustering coefficient of a random graph with the same number of nodes and edges.

They propose several models of graphs, one of which is called β-graphs. Starting from a regular graph, they introduce disorder into the graph by randomly rewiring each edge with probability p as shown in Fig.3.1. If $p = 0$, then the graph is completely regular and ordered. If $p = 1$ then the graph is completely random and disordered. Intermediate values of p give graphs that are neither completely regular nor completely disordered. They are small worlds.

For example, Fig.3.2 is a graph constructed from a document as follows[1]: first the document is preprocessed by stemming and removing *stop words* as in [3.20], and extracting an n-gram. Then, each sentence of the document is considered to be in a basket, each of which consists of words (or phrases). After the preprocess, nodes are settled by selecting a word which appears over a user-given threshold number of times (e.g. three times). For every pair of nodes, the co-occurrence for every sentence is counted; an edge is added if the Jaccard coefficient exceeds a threshold, $J_{\rm thre}$. The Jaccard coefficient is simply the number of sentences that contain both terms divided by the number of sentences that contain either term. This idea is also used in constructing a referral network from WWW pages [3.13]. Figure 3.2 shows a graphical visualization of the world of a document. Nodes are clustered, yet the whole graph is connected loosely. The co-occurrence graph of a technical paper comprises a small world.

Recently, many studies have revealed small-world characteristics. Mathias and Gopal investigated small-world networks from the point of view of their origin

[1] Note that *KeyGraph* was also invented as a document-processing algorithm.

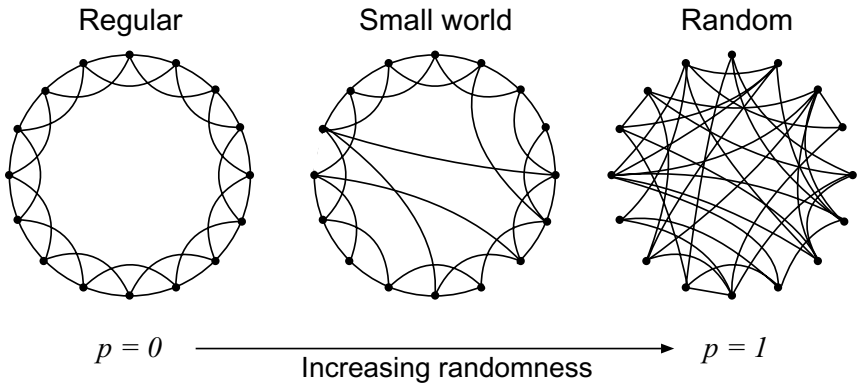

Fig. 3.1. Random rewiring of a regular ring lattice

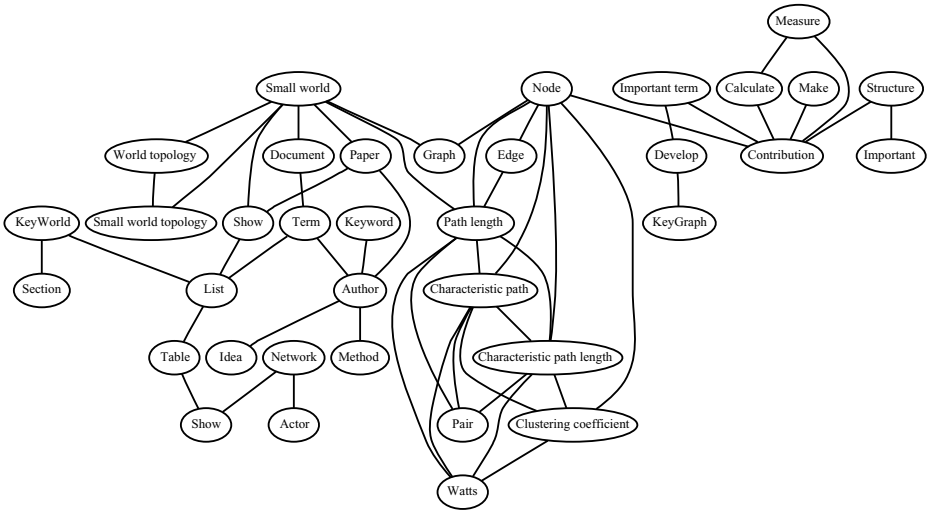

Fig. 3.2. Small world of a document

[3.16]. They showed that small-world topology arises as a consequence of a tradeoff between maximal connectivity and minimal wiring.

3.4.2 Structural Importance

In [3.18], node contribution is considered in the context of a small-world: if a node is to be deleted, at what point will the small-world topology break? The contribution of node v, CB_v, is measured by

$$\mathrm{CB}_v = L_{G_v} - L_v, \tag{3.3}$$

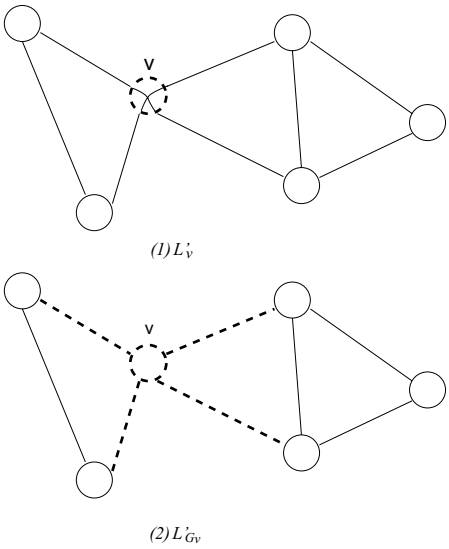

Fig. 3.3. L_v and L_{G_v}

where L_v is the characteristic path length averaged over all pairs of nodes except node v and L_{G_v} is the characteristic path length of the graph without node v 3.3.

The larger CB_v is, the greater its contribution to a small world. We can detect which nodes are structurally important from the viewpoint of a small world, that is, those that contribute to the efficiency of network flow and efficiency of network cost.

This method is based on the same idea as *KeyGraph*: if a node (an event) shares an important position in a graph, it might have an impact even if the frequency of the event is low. Importance is defined, in the *KeyGraph* case, by co-occurrence of two or more big clusters, and in the small-world case by the contribution for the graph to be highly connected. The method in [3.17] employs another importance criterion: if the flow on the graph is through a certain node, the node is important.

There can be many other ways to define importance on the network. However, this direction seems promising because the structure (or context) is considered to evaluate the importance of events. Certainly, such methods will not detect important rare events by themselves, but by being used in combination with human understanding, they have great potential for data analysis and prediction (or even invention) of the future.

3.5 Conclusion

This chapter gives an overview of prediction methods including the ARMA model, pattern recognition, and data mining; differences and relevance between prediction

and forecasting and chance discovery are discussed. Although the presuppositions of prediction and forecasting and chance discovery differ, both will be useful in different stages of data analysis and decision making.

References

3.1 Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In Bocca,J.B, Jarke,M, and Zaniolo, C, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, Morgan Kaufmann, San Francisco, CA, pp. 487–499

3.2 Agrawal R, Srikant R (1995) Mining sequential patterns. In Philip S. Yu and Arbee S. P. Chen, editors, *Proceedings of the Eleventh International Conference on Data Engineering*, IEEE Computer Society Press, Los Alamoitos, CA pp. 3–14,

3.3 Albert A, Jeong H, Barabási A (1999) Diameter of the World-Wide Web. *Nature* 401(6749)

3.4 Bialek W, Nemenman I, and Tishby N (2001) Predictability, complexity, and learning. *Neural Computation*, 13:2409–2463

3.5 Brockwell PJ, Davis R (1996) *Introduction to Time-Series and Forecasting*. Springer Verlag, Heldelberg, Germany

3.6 Domeniconi C, Perng C, Vilalta R, and Ma S (2002) A classification approach for prediction of target events in temporal sequences. In *Proc. 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02)*, pp.125 – 137

3.7 Chatfield C (1996) *The Analysis of Time Series*. Chapman and Hall, London, UK, 5th edition

3.8 Embrechts P, Klüppelberg C, and Mikosch T (1991) *Modelling Extremal Events for Insurance and Finance*. Springer Verlag, Heidelberg, Germany

3.9 Garofalakis MN, Rastogi R, and Shim K (1999) SPIRIT: Sequential pattern mining with regular expression constraints, In *Proc. 25th International Conference on Very Large Data Bases (VLDB'99)*, pp.223–234

3.10 Geurts P (2001) Pattern extraction for time series classification. In *Proc. PKDD 2001*, pp.115–127

3.11 Gladwell M (2000) *The Tipping Point: How Little Things Can Make a Big Difference*. Little Brown & Co, Boston, MA

3.12 Han J, Gong W, Yin Y (1998) Mining segment-wise periodic patterns in time-related databases. In *Fourth International Conference on Knowledge Discovery and Data Mining*, AAAI Press, Menlo Park, CA, pp.214–218

3.13 Kautz H, Selman B, Shah M (1997) The hidden Web. *AI magazine*, 18(2):27–35

3.14 Last M, Klein Y, Kandel A (2001) Knowledge discovery in time series databases. *IEEE Transactions on Systems, Man, and Cybernetics*, 31(1): 160 – 169

3.15 Mannila H, Toivonen H, Verkamo AI (1995) Discovering frequent episodes in sequences. In *Proc. 1st International Conference on Knowledge Discovery and Data Mining (KDD'95)* pp.210 – 215

3.16 Mathias N, Gopal V (2001) Small worlds: How and why. *Physical Review E*, 63(2):021117 – 021128

3.17 Matsumura N, Ohsawa Y, Ishizuka M (2002) Pai: Automatic indexing for extracting asserted keywords from a document. In *Proc. AAAI Fall Symposium on Chance Discovery* pp.28 – 32

3.18 Matsuo Y, Ohsawa Y, Ishizuka M (2001) A document as a small world. In *Proceedings the 5th World Multi-Conference on Systemics, Cybenertics and Informatics (SCI2001)*, 8: 410–414

3.19 Milgram S (1967) The small-world problem, *Psychology Today*, 2:60–67

3.20 Salton G (1989) *Automatic Text Processing*. Addison-Wesley, Boston, MA
3.21 Shannon CE (1948) A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656
3.22 Singh S (2000) Pattern modelling in time-series forecasting. *Cybernetics and Systems - An International Journal*, 31(1): 49 – 66
3.23 Sornette D (2002) Predictability of catastrophic events: material rupture, earthquakes, turbulence, financial crashes and human birth. In *Proc. National Academy of Sciences USA* pp.60 – 67
3.24 Srikant R, Agrawal R (1996) Mining sequential patterns: Generalizations and performance improvements. In Peter M. G. Apers, Mokrane Bouzeghoub, and Georges Gardarin, editors, *Proc. 5th Int. Conf. Extending Database Technology, EDBT*, volume 1057, pp.3–17, Springer Verlag, Heidelberg, Germany
3.25 Tan N, Kumar V (2001) Mining indirect associations in web data. *Proc. of WebKDD 2001: Mining Log Data Across All Customer TouchPoints*, pp.145 – 166
3.26 Tan PN, Kumar V, Srivastava J (2000) Indirect association: Mining higher order dependencies in data. In *Proc. the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp.632–637
3.27 Watts D (1999) *Small worlds: the dynamics of networks between order and randomness*. Princeton University Press, Princeton, NJ
3.28 Weigend AS, Gershenfeld NA (1993) *Time Series Prediction*. Addison-Wesley, Boston, MA
3.29 Weiss GM, Hirsh H (1998) Learning to predict rare events in event sequences. In R. Agrawal, P. Stolorz, and G. Piatetsky-Shapiro, editors, *Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*, New York, NY, 1998. AAAI Press, Menlo Park, CA, pp.359–363
3.30 Weiss GM (1999) Timeweaver: a genetic algorithm for identifying predictive patterns in sequences of events. In *Proc. the Genetic and Evolutionary Computation Conference (GECCO-99)*, pp.718–725
3.31 Yule GU (1927) On a method of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers. Philosophical Transactions of the Royal Society of London, Series A, 226:267–298