A SURVEY OF LINK MINING AND ANOMALIES DETECTION

Dr. Zakea Idris Ali

Staffordshire University- UK

ABSTRACT

This survey introduces the emergence of link mining and its relevant application to detect anomalies which can include events that are unusual, out of the ordinary or rare, unexpected behaviour, or outliers.

KEYWORDS

Link mining, anomalies detection.

1. INTRODUCTION

Link mining is a new emerging research area, which differs from data mining. Whilst data mining aims at discovering new potentially hidden patterns in datasets, link mining considers datasets as a linked collection of interrelated objects and therefore it focuses on discovering explicit links between objects. A crucial step in both data and link mining is to ensure that the analysis is undertaken on reliable, robust and efficient data, and to identify outliers, which are observations that are numerically distant from the rest of the data. Reliability of detection anomaly should achieve high data delivery reliability unless the quality of the underlying links makes that infeasible. Robustness should be robust against huge or complex social networks failures, dynamic networks, and topology changes. In spite of these dynamics, it should function without much tuning or configuration. Efficiency in communication often applies both complex anomalies and different types of anomalies, to allow an opportunity to make the method detection anomalies more efficient. Though outliers are often considered as an error or noise in data mining, they are often referred to as anomalies in link mining as they can carry important information. Often the data contains noise that tends to be similar to the actual anomalies and hence it is difficult to distinguish and remove them (Chandola et al., 2009). Any errors in data are to be examined taking into consideration the context of the domains; some may be true errors and therefore removed, whereas other errors may be regarded as interesting anomalies.

Link mining applications have been shown to be highly effective in addressing many important business issues such as money laundering (Kirkland et al., 1999), telephone fraud detection (Fawcett and Provost 1999), crime detection (Sparrow 1991), terrorism (Badia and Kantardzic 2005, Skillicorn 2004), the financial domain (Creamer and Stolfo 2009), social networks and health care problems (Provana et al., 2010, Wadhah et al., 2011). The identification of anomalies is affected by various factors, many of which are of interest for practical applications. For example, criminal deception or fraud will constantly be a costly issue for many profit

DOI : 10.5121/ijdkp.2016.6201

organisations. Link mining can minimise some of these losses by making use of the massive collections of customer data (Phua et al., 2004) Using web log files, it becomes possible to recognise fraudulent behaviour, changes in behaviour of customers, or faults in systems. Anomalies arise by reasons of such incidents. Consequently, typical fault detection can discover exceptions in the type of items purchased, the amount of money spent, the time and the location of this purchase information such as the name of the credit holder account number and expiry date which are very easy to obtain, even from one's home mailbox or from any online transaction carried out (Alfuraih et al., 2002). Such automatic systems aimed at preventing fraudulent use of credit cards; detecting unusual transactions are therefore desirable.

Knowledge discovery is the non-trivial removal of implicit, previously unknown, and potentially useful information from data. The type of knowledge that is discovered from databases and its corresponding representational form varies widely depending on both the application area and the database type, such as data mining, text mining, web mining and link mining. The specification of the type of knowledge to be discovered directs the pattern-filtering process. Data mining involves the use of complicated data analysis tools to discover previously unknown, relationships and valid patterns in large data sets. These tools involve mathematical algorithms, machine-learning methods and statistical models, and applications such as banking, insurance and medicine; while text mining has been applied to semi-structured and unstructured information, such as digital libraries and biological information systems. Technologies in the text-mining process include information extraction, topic tracking, summarisation, categorisation, clustering, and concept linkage information extraction (Chakrabarti, 2001). Web mining is the extraction of interesting and potentially useful patterns and implicit information from activity related to the World Wide Web whereas link mining, focuses on discovering explicit links between objects.

Anomalies detection, which is the focus of this paper, is concerned with the problem of finding non-conforming patterns in data sets, such as social network, bibliometrics data and citation. Anomalies can include exceptions, outliers, aberrations, surprises, peculiarities, and so on (Chandola et al., 2009). In data, text and link mining, the first task is to pre-process the data to explore their integrity. Any errors observed in the data, must be analysed within the context of domains and purpose of the analysis.

2. EMERGENCE OF LINK MINING

Link mining attempts to build predictive or descriptive models of the linked data (Getoor & Diehl, 2005). The term 'link' in the database community differs from that in the AI community. In this research a link refers to some real-world connection between two entities (Senator, 2005). Link mining focuses on techniques that explicitly consider these links when building predictive or descriptive models of the data sets (Getoor, 2005). In data mining, the main challenge is to tackle the problem of mining richly structured heterogeneous data sets. The data domains often consist of a variety of object types; these objects can be linked in a variety of ways. Traditional statistical inference procedures assume that instances are independent and this can lead to unsuitable conclusions about the data. However, in link mining, object linkage is a knowledge that should be exploited. In many applications, the facts to be analysed are dynamic, so it is important to develop incremental link mining algorithms, besides mining knowledge from link objects and networks (Getoor & Diehl, 2005).

3. LINK MINING TASKS

In their paper, Getoor and Diehl (2005) identify a set of link mining tasks (see Figure 1), which are:

- Object-related tasks.
- Graph-related tasks.
- Link-related tasks.

3.1 Object-related tasks

These tasks include link-based object clustering, link-based object classification, object identification and object ranking. In a bibliographic domain, the objects include papers, authors, institutions, journals and conferences. Links include the paper citations, authorship and co-authorship, affiliations, and the relation between a paper and a journal or conference.

3.2 Graph-related tasks

These tasks consist of sub-graph discovery, graph classification, and generative models for graphs. The aim is to cluster the nodes in the graph into groups sharing common characteristics. In the bibliographic domain, an example of graph classification is predicting the category of a paper, from its citations, the papers that cite it, and co-citations (papers that are cited with this paper).

3.3 Link-related tasks

These tasks aim at predicting the existence of a link between two entities based on the attributes of the objects and other observed links. In a bibliographic domain, predicting the number of citations of a paper is an indication of the impact of a paper— papers with more citations are more likely to be seminal.

Link prediction is defined as inferring the existence of a link (relationship) in the graph that is not previously known. Examples include predicting links among actors in social networks, such as predicting friendships or predicting the participation of actors in events (O'Madadhain et al., 2005) such as email, telephone calls and co-authorship. Some links can be observed, but one is attempting to predict unobserved links, or monitor the temporal aspect; for example, if a snapshot of the set of links at time t is observed then the goal is to predict the links at time t + 1.

This problem is normally expressed in terms of a simple binary classification problem. Given two potentially linked objects Oi and Oj, the task is to predict whether Lij is 1 or 0. One approach bases the prediction on the structural properties of the network, for example using predictors based on different graph proximity measures Liben-Nowell and Kleinberg (2003). The second approach is to use attribute information to predict a link. Popescul et al. (2003) applied a structured logistic regression model on relational features to predict the existence of links. A conditional probability model is proposed which is based on attribute and structural features by O'Madadhain et al (2005); (Getoor, 2003; O'Madadhain, 2005; Rattigan & Jensen, 2005). They explain that building statistical models for edge prediction is a challenging problem because the

prior probability of a link can be quite small, this makes it difficult to evaluate the model and, more importantly, measure the level of confidence in the predictions. Rattigan and Jensen (2005) propose improving the quality of the predictions by making the predictions collectively. Hence, a number of probabilistic approaches have been developed, some network structure models are based on the Markov Random Field (MRF) model (Chellappa & Jain, 1993) others on Relational Markov Network (Taskar et al., 2003) and, more recently, the Markov Logic Network (Domingos & Richardson, 2004). If case, O represents a set of objects, with X attributes, and E edges among the objects, then MRF uses a joint distribution over the set of edges E, P(E), or a distribution conditioned on the attributes of the nodes, P(E/X). Getoor et al (2003) described several approaches for handling link uncertainty in probabilistic relational models. The key feature of these approaches is their ability to perform probabilistic inferences about the links, which allows the capture of the correlations among the links. This approach is also used for other tasks, such as link-based classification, which allow for more accurate predictions. Hence, approximate inference techniques are necessary to join the model-based probabilistic approaches based on their computational cost to exact inference as general intractable goals.

Desjardins and Gaston (2006) discuss the relationship between the fields of statistical relational learning (SRL) and multi-agent systems (MAS) using link prediction methods to recognise collusion among agents, and applying graph classification to discover efficient networks for MAS problems. Mustafa et al. (2007) show a general approach for combining object classification and link prediction using Iterative Collective Classification and Link Prediction (ICCLP) in graphs.

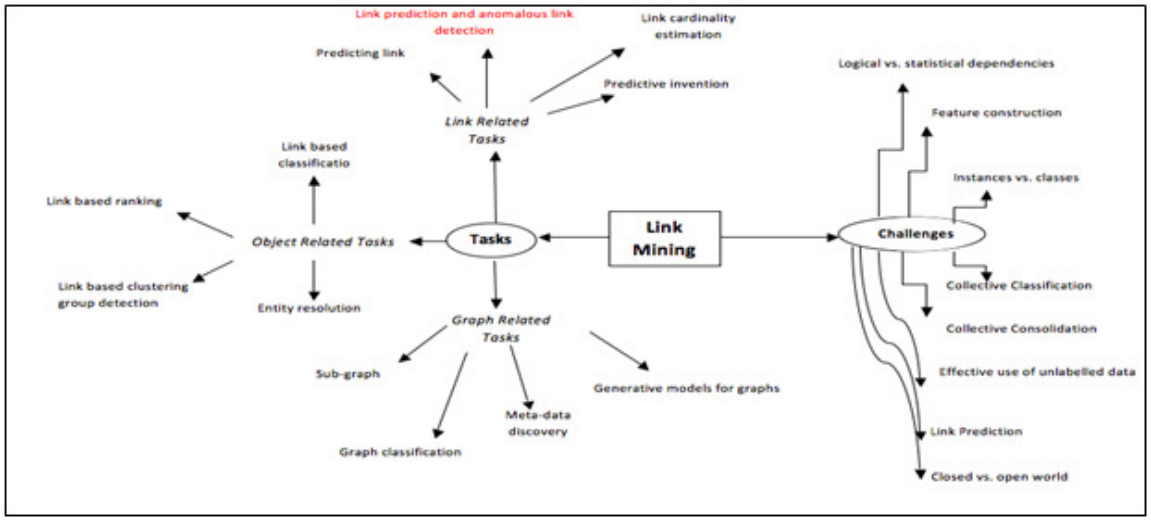

Figure 1. Link mining tasks and challenges

4. LINK MINING CHALLENGES

Research into link mining involves a set of challenges associated with these tasks, as Senator (2005), Getoor (2005) and Pedreschi (2008) explain (see Figure 1). These are:

- logical vs statistical dependencies that relate to the identification of logical relationships between objects and statistical relationships between the attributes of objects;

- feature construction, which refers to the potential use of the attributes of linked objects;
- collective classification using a learned link-based model that specifies a distribution over link and content attributes, which may be correlated through these links;
- effective use of unlabelled data using semi-supervised learning, co-training and transductive inference to improve classification performance;
- link prediction, which predicts the existence of links between objects;
- object identity, that is, determining whether two objects refer to the same entity; and closed world vs open world assumptions of whether we know all the potential entities in the domain.
- the challenge of this study is to identify and interpret anomalies among the observed links.

5. APPLICATIONS OF LINK MINING

An application for each of the three tasks is listed below.

- Social bookmarking is an application of a link-related task. Tools enable users to save URLs for upcoming reference, to create labels for annotating web pages, and to share web pages they found interesting with others. The application of link mining to social web bookmarking investigates user bookmarking and tagging behaviours, and describes several approaches to finding patterns in the data (Chen & Pang-Ning, 2008).

- Epidemiological studies are an application associated with object-related task. In an epidemiology domain, the objects include patients, people with whom they have come into contact and disease strains. Links represent contacts between people and a disease strain with which a person is infected (Getoor, 2003).

- Friendship in a social network is an application of graph-related task. This is annotated by the inclusion of the friend's name on a user's homepage. Pair-dependent descriptions, such as the size of the intersection of interests, offer supplementary evidence for the existence of a friendship. These pair-dependent features are used to determine the probability for link existence where it is not annotated. Finding the non-obvious pair-dependent features can be quite difficult as it, requires the use of recent developments in association rule mining and frequent pattern mining to find correlations between data points that best suggest link existence (Han *et al*., 2001).

- Bibliographic area is an application of a graph-related task. Information networks are mainly new. Link information in a bibliographic database provides in-depth information about research, such as the clustering of conferences shared by many common authors, the reputation of a conference for its productive authors, research evolving with time, and the profile of a conference, an author, or a research area. This motivates the study of information network in link mining on bibliographic databases (Getoor, 2003).

- Discovery of a fundamental organisation is an application of graph-related task. Structure from crime data leads the investigation to terrorist cells or organised crime groups, detecting covert networks that are important to crime investigation. (Marcus *et al*., 2007).

6. ANOMALIES DETECTION

Link prediction is a complex and challenging task as many applications contain data which are extremely noisy and often the characteristics to be employed for prediction are either not readily available or involve complex relationships among objects. The focus of this paper is to investigate the links between objects and understand the context of their anomalies. Anomaly detection is different from noisy data, which is not of interest to the analyst, and must be removed before any data analysis can be performed. In our research anomalous objects or links can convey useful information and should be investigated.

Song *et al*. (2007) and Chandola *et al.* (2009) describe five types of anomalies, these are:

- Contextual anomalies (also known as conditional anomalies) refer to data instances anomalous in a specific context. A temperature of 5^{o}C might be normal during the winter period in the UK, but would be an anomaly in the summer time.

- Point anomalies refer to a data instance anomalous with respect to the rest of the data set. In credit card fraud application, a transaction is considered a point anomaly if it contains a very high amount spent compared to the normal range of expenditure for that individual.

- Collective anomalies refer to a set of data instances anomalous with respect to the entire data set. For example an electrocardiogram output may show a region of low values for an abnormally long time due to some premature contractions (Goldberger *et al.,* 2000). These low values may not be anomalies by themselves, but their existence together as a collection is anomalous.

- On-line anomalies refer to data present often in a streaming mode where the normal behaviour is changing dynamically.

- Distributed anomalies refer to detecting anomalies in complex systems.

The definition of anomaly is dependent on the type of application domains. For example, in the medical domain a small deviation from normal (e.g., fluctuations in body temperature) could be an anomaly, however similar deviation in the stock market domain (e.g., fluctuations in the value of a stock) might be considered as normal. Thus applying a technique developed in one domain to another has to take into consideration the context of that domain.

Anomalies detection is alike to link prediction in the sense that they both use similar metrics to evaluate which links are anomalous and which ones are expected. Thus research on improving either problem should benefit the other. Rattigan and Jensen explain that one of the important challenges in link prediction is to address the problem of a highly skewed class distribution caused by the fact that " ... as networks grow and evolve, the number of negative examples (disconnected pairs of objects) increases quadratically while the number of positive examples often grows only linearly" (Rattigan and Jenssen 2005: 41). As a result, evaluating a link

prediction model becomes a complex task and computationally costly because of the need to evaluate all potential links between all pairs of objects. They have proposed the alternative task of anomalous link discovery (ALD) focusing on those links that are anomalous, statistically unlikely, and most "interesting" links in the data. Typical applications of anomaly detection algorithms are employed in domains that deal with security and privacy issues, terrorism activities, picking intrusion detection and illegitimate financial transactions (See Figure 1).

7. ANOMALIES DETECTION APPROACHES AND METHODS

A survey of the literature reveals three main approaches used to detect anomalies. These are described below:

- *Supervised* anomalies detection operates in supervised mode and assumes the availability of a training data set, which has labels available for both normal and anomalous data. Typical approach in such cases is to build a predictive model for normal vs. anomalous classes; their disadvantage is that they require labels for both normal and anomalous behaviour. Certain techniques insert artificial anomalies in a normal data set to obtain a fully labelled training data set and then apply supervised anomalies detection techniques to detect anomalies in test data (Abe *et al*., 2006).

- *Semi-supervised* anomalies detection, which models only normality and are more applicable than the previous approach since only labels for normal data is required. Such techniques are not used commonly, as it is difficult to obtain a training data set, which covers possible outlying behaviour that can occur in the data (Chandola *et al.,* 2009).

- *Unsupervised* anomalies detection, which makes the implicit assumption that normal instances are more frequent than anomalies in the test data. If this assumption is not true then such techniques suffer from a high false alarm rate (Chandola *et al.,* 2009).

Unsupervised method is very useful for two reasons. First, they do not rely on the availability of expensive and difficult to obtain data labels; second, they do not assume any specific characteristics of the anomalies. In many cases, it is important to detect unexpected or unexplained behaviour that cannot be pre-specified. Since the unsupervised approach relies on detecting any observation that deviates from the normal data cases, it is not restricted to any particular type of anomaly.

In their paper, Chandola *et al. (*2009) identify five different methods employed in anomalies detection: nearest neighbour, clustering, statistical, classification, and information/ context based approaches (see Figure 2).

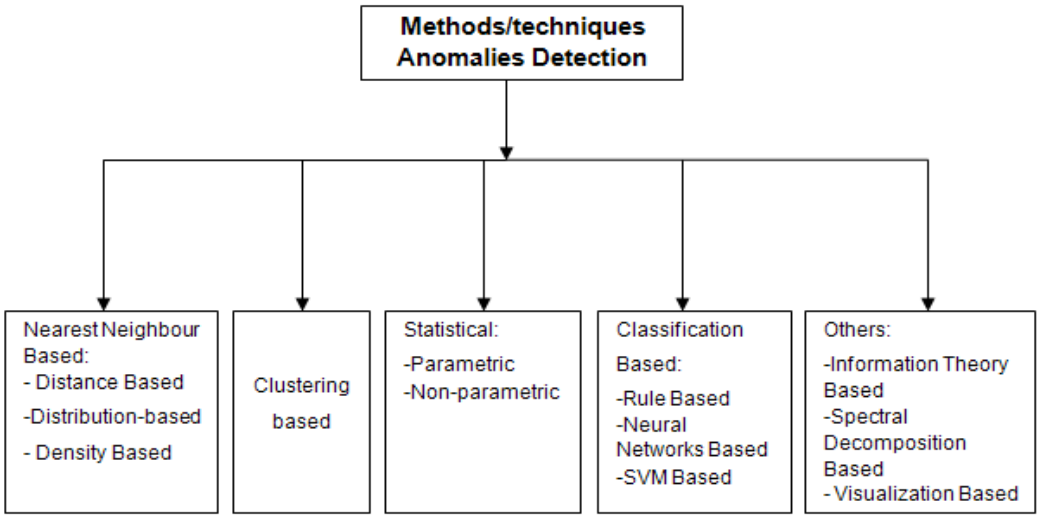

Figure 2. Methods of anomalies detection

7.1 Nearest neighbour based detection techniques

The concept of nearest neighbour has been used in several anomaly detection techniques. Such techniques are based on the following key assumption:

***Assumption**:* Normal data instances happen in dense neighbourhoods, while anomalies occur far from their closest neighbours.

The nearest neighbour based method can be divided into three main categories. The first distance-based methods, distinguish potential anomalies from others based on the number of objects in the neighbourhood (Hu and Sung, 2003). The distribution-based approach deals with statistical methods that are based on the probabilistic data model, which can be either a automatically or priori, created using given data. If the object does not suit the probabilistic model, it is considered to be an outlier (Petrovskiy, 2003). The density-based approach detects local anomalies based on the local density of an object's neighbourhood (Jin *et al*., 2001). A typical application area is fraud detection (Ertoz *et al.*, 2004; Chandola et al. 2006), Eskin *et al* (2002).

Nearest neighbour based techniques have many advantages. Key advantage is that they are unsupervised in nature and do not make any assumptions concerning the generative distribution of the data. Instead, it is purely data driven. Adapting these techniques to a variety of data type requires defining a distance measure for the given data. With regards to mixed anomalies, semi-supervised techniques perform more improved than unsupervised techniques since the likelihood of an anomaly is to form a near neighbourhood when the training data set is low.

However, these techniques have disadvantages. They fail to label the anomalies correctly, resulting in missed anomalies, for unsupervised techniques. If the data has normal instances that do not have close neighbours or if the data has anomalies that have close neighbours the technique fails to label them correctly, resulting in missed anomalies. The computational complexity of the testing phase is a challenge since it involves computing the distance of each test instance with all instances belonging to either the test data itself, or to the training data. In

semi-supervised techniques, if the normal instances in the test data do not have enough similar normal instances in the training data, then the technique will have a high false positive rate.

7.2 Clustering-based anomalies detection techniques

Clustering-based anomalies detection techniques can be grouped into three assumptions:

The first assumption: *Normal data instances belong to a cluster in the data, while anomalies do not belong to any cluster.* Techniques based on this assumption apply a known clustering-based algorithm to the data set and declare any data instance that does not belong to any cluster as anomalous. Several clustering algorithms do not force every data instance to belong to a cluster, such as *DBSCAN* (Ester *et al*., 1996), *ROCK* (Guha *et al.,* 2000) and *SNN clustering* (ErtÄoz *et a*l., 2003). The *FindOut* algorithm (Yu *et al*., 2002) is an extension of the *WaveCluster* algorithm (Sheik-holeslami *et al.,* 1998) in which the detected clusters are removed from the data and the residual instances are declared as anomalies. A disadvantage of these techniques is that they are not optimised to find anomalies, as the main aim of the underlying clustering algorithm is to find clusters. Typical application areas include image processing (Scarth *et a*l., 1995), and fraud detection (Wu and Zhang, 2003; Otey *et a*l. 2003).

The second assumption: *Normal data instances lie close to their closest cluster centroid, while anomalies are far away from their closest cluster centroid.* Techniques based on this assumption consist of two steps. In the first step, the data is clustered using a clustering algorithm. In the second step, for each data instance, its distance to its closest cluster centroid is calculated as its anomaly score. A number of anomaly detection techniques that follow this two-step approach have been proposed using different clustering algorithms. Smith *et al*. (2002) study *Self-Organizing Maps (SOM)*, *K-means* and *Expectation Maximization (EM)* to cluster training data and then use the clusters to classify test data. In particular, SOM (Kohonen, 1997) has been widely used to detect anomalies in a semi-supervised mode in several applications such as intrusion detection (Labib and Vemuri, 2002; Smith *et al*., 2002; Ramadas *et al*., 2003), fault detection (Harris, 1993; Ypma, Duin, 1998; Emamian *et al*., 2000) and fraud detection (Brockett *et al.,* 1998). Barbara *et al*. (2003) propose a robust technique to detect anomalies in the training data. This assumption can also operate in a semi-supervised mode, in which the training data are clustered, with instances belonging to the test data being compared against the clusters to obtain an anomaly score for the test data instance (Marchette, 1999; Wu and Zhang, 2003; Vinueza & Grudic, 2004; Allan *et al*., 1998). If the training data have instances belonging to multiple classes, semi-supervised clustering can be applied to improve the clusters to address this issue.

The third assumption: *Normal data instances belong to large and dense clusters, while anomalies belong either too small or too sparse clusters.* Techniques based on the above assumption declare instances belonging to cluster as anomalous if size/density is below a threshold. Several variations of the third assumption of techniques have been proposed (Pires and Santos-Pereira, 2005; Otey *et al*., 2003; Eskin *et al*., 2002; Mahoney *et al*., 2003; Jiang *et al*., 2001; He *et al*., 2003). The technique proposed by He *et al*. (2003), called *FindCBLOF*, assigns an anomaly score known as the Cluster-Based Local Outlier Factor (CBLOF) to each data instance. The CBLOF score captures the size of the cluster to which the data instance belongs, in addition to the distance of the data instance to its cluster centroid. These techniques are used for network intrusion detection (Bolton & Hand 1999), and for host based intrusion detection (Sequeira & Zaki 2002).

In terms of advantages these techniques can work in an unsupervised mode, and can be adapted to complex data types by working in a clustering algorithm that can handle the specific data type. The testing stage for clustering based techniques is fast because the number of clusters against is a small constant. However these techniques are highly dependent on the effectiveness in capturing the cluster structure of normal instances. Numerous techniques detect anomalies as a result of clustering, and are not improved for anomaly detection. Some clustering algorithms are assigned to a particular cluster. This could result in anomalies getting assigned to a larger cluster, thus being considered as normal instances by techniques that work under the assumption that anomalies are not linked to any cluster. If O (N2d) clustering algorithms are used, then the computational complexity for clustering the data is often a bottleneck.

7.3 Statistical techniques

Statistical anomaly detection techniques are based on the following key assumption: ***Assumption****:* Normal data instances occur in high probability regions of a stochastic model, while anomalies occur in the low probability regions of the stochastic model.

Statistical techniques operate in two phases: *training* and *testing* phases, once the probabilistic model is known. In the *training* phase, the first step comprises fitting a statistical model to the given data, whereas the *testing* phase, determines whether a given data instance is anomalous with respect to the model or not. This involves computing the probability of the test instance to be generated by the learnt model. Both parametric and non-parametric techniques are used. Parametric techniques assume the knowledge of underlying distribution and estimate the parameters from the given data (Eskin 2000). Non-parametric techniques do not assume any knowledge of distribution characteristics (Desforges *et al*., 1998). Typically the modelling techniques are robust to small amounts of anomalies in the data and hence can work in an unsupervised mode. Statistical techniques can operate in unsupervised settings, semi-supervised and supervised settings. Supervised techniques estimate the probability density for normal instances and outliers. The semi-supervised techniques estimate the probability density for either normal instances, or anomalies, depending on the availability of labels. Unsupervised techniques define a statistical model, which fits the majority of the observations. One such approach is to find the distance of the data instance from the estimated mean and declare any point above a threshold to be anomalies (Grubbs 1969). This requires a threshold parameter to determine the length of the tail, which has to be considered as anomalies; techniques used for mobile phone fraud detection (Cox *et al*., 1997).

The advantages of these techniques are as follows:

- If the assumptions concerning the underlying data distribution are true, these techniques then offer a statistically correct solution for anomaly detection.

- Confidence interval is associated with the anomaly score provided by a statistical technique, which can be used as extra information when making a decision concerning any test instance.

- It can operate in an unsupervised setting without any need for labelled training data if the distribution estimation step is robust to anomalies in data.

However, they rely on the assumption that the data is conducted from a particular distribution. This assumption is not necessarily true, particularly for high dimensional real data sets. Even when the statistical assumption can be justified, there are several hypothesis test statistics that can be useful to detect anomalies; choosing the greatest statistic is often not an easy task (Motulsky 1995). In specific, composing hypothesis tests for complex distributions needed to fit high dimensional data sets is nontrivial. An anomaly might have attribute values that are individually very common, but their combination is very uncommon, but an attribute-wise histogram based technique would not be able to detect such anomalies. Histogram based techniques are relatively simple to apply, a key disadvantage of such techniques with regards to multivariate data is that they are not able to capture the interactions between different attributes.

7.4 Classification techniques

Classification based techniques operate under the following general assumption:

Assumption: A classifier that can distinguish between normal and anomalous classes can be learnt in the given feature space.

Classification is an important data-mining concept. The aim of classification is to learn a set of labelled data instances (training) and then classify an unseen instance into one of the learnt class (testing). Anomalies detection techniques based on classification also operate in the same two-phase, using normal and anomalies as the two classes. The training phase builds a classification model using the available labelled training data. The testing stage classifies a test instance using the model learnt. The techniques following this approach fall under supervised anomalies detection techniques. A one-class classifier can then be trained to reject this object and to label it as anomalies. These techniques fall under the category of semi-supervised anomalies detection techniques (Tan *et al*. 2005b; Duda *et al*. 2000).

The classification problem is modelled as a two-class problem where any new instance that does not belong to the learnt class is anomalous. In real scenarios, class labels for normal class are more readily available but there are also cases where only anomalies class labels are available. Classification based techniques are categorised into subcategories based on the type of classification model that use. These include Neural networks, Bayesian Networks, Support Vector Machines (SVM), decision trees and regression models. These rules are used to classify a new observation as normal or anomalous. In term of advantages, the testing stage of these techniques is fast since each test instance needs to be compared against the pre-computed model. They can make use of powerful algorithms that can differentiate between instances belonging to different classes. However, Multi-class classification techniques rely on availability of precise labels for different normal classes, which is often not possible. These techniques allocate a label to each test instance, which can become a disadvantage when a meaningful anomaly score is wanted for the test instances. Some classification techniques that obtain a probabilistic prediction score from the output of a classifier can be used to address this issue (Platt 2000).

7.5 Information Theory Based

These techniques are based on the following key assumption:

Assumption: Anomalies in data induce irregularities in the information content of the data set.

Information theory based techniques analyse the information content of a dataset using different information theoretic measures such as relative entropy, entropy, *etc*. The general idea is that normal data is regular in terms of a certain information theoretic measure. Anomalies significantly change the information content of the data because of their surprising nature. Thus, the typical approach adopted by this technique is to detect data instances that induce irregularity in the data, where the regularity is measured using a particular information theoretic measure. Information theory based techniques operate in an unsupervised mode.

The advantages of these techniques are as follows:

- They can function in an unsupervised setting.
- They make no assumptions regarding the underlying statistical distribution of the data.

However, the performance of these techniques is greatly dependent on the choice of the information theoretic measure. Frequently, these measures can detect anomalies only when there are large numbers of anomalies existing in the data. It is often nontrivial to obtain when these techniques are applied to sequences and spatial data sets because they rely on the size of the substructure. Another disadvantage is that it is difficult to associate an anomaly score with a test instance using these techniques.

7.6 Other Techniques

These techniques are based on the following key assumption:

Assumption: Data can be embedded into a lower dimensional subspace in which normal instances and anomalies appear significantly different.

Spectral decomposition based technique finds an approximation of the data using a combination of attributes that capture the size of variability in the data. The underlying assumption for such techniques is that the reduced sets of attributes faithfully capture much of the normal data, but this is not necessarily true for the anomalies. Spectral techniques can work in an unsupervised as well as semi-supervised setting. This approach has been applied to the network intrusion detection domain by several different groups (Shyu *et al*. 2003; Lakhina *et al.* 2005; Thottan and Ji 2003) and for detecting anomalies, for example in spacecraft components (Fujimaki *et al*. 2005).

Visualisation based technique maps the data in a coordinate space that makes it easy to visually identify the anomalies. Cox *et al.* (1997) present a visualisation-based technique to detect telecommunications fraud, which displays the call patterns of various users as a directed graph such that a user can visually identify abnormal activity.

These techniques routinely perform dimensionality reduction, which makes them suitable for handling high dimensional data sets. Additionally, they can be used as a pre-processing step, followed by application of any existing anomaly detection technique in the transformed space. These techniques can be used in an unsupervised setting.

However, these techniques usually have high computational complexity. They are useful only if normal and anomalous instances are separate in the lower dimensional embedding of the data.

7.7 Overview of strengths and limitations

For high-dimensional data, any of the above anomalies detection techniques can easily detect the anomalies. For more complex data sets, different techniques face different challenges. Chandola *et al*. (2009) argue that statistical techniques do not work well with high-dimensional categorical data and that visualisation-based techniques are more naturally suited to low-dimensional data and hence require dimensionality reduction as a pre-processing step when dealing with a higher number of dimensions. Spectral decomposition-based techniques, which find an approximation of the data using a combination of attributes to capture the variability in the data, explicitly address the high-dimensionality problem by mapping data to a lower dimensional projection, but their performance is highly dependent on the fact that the normal instances and anomalies are distinguishable in the projected space. Clustering is often called an unsupervised learning task, as no class values indicate an a priori grouping of the data instances, as in the case for supervised learning. Clustering and nearest neighbour techniques rely on a good similarity or distance measure to handle the anomalies in complex data sets. Classification-based techniques handle the dimensionality better, since they try to assign weights to each dimension and ignore unnecessary dimensions automatically. However, classification-based techniques require labels for both normal data and anomalies. Finally, information theory-based techniques, which analyse the information content of a data set using different information theoretic measures (e.g. entropy measure), require a measure that is sensitive enough to detect the effects of even single anomalies. Such techniques detect anomalies only when there is a significant number of an anomaly.

8. CHALLENGES OF ANOMALIES DETECTION

Multi- and high-dimensional data make the outlier mining problem more complex because of the impact of the curse of dimensionality on algorithms' performance and effectiveness. Wei *et al*., (2003) introduce an anomalies mining method based on a hyper-graph model to detect anomalies in a categorical data set. He *et al*. (2005) define the problem of anomalies detection in categorical data as an optimisation problem from a global viewpoint, and present a local search heuristic-based algorithm for efficiently finding feasible solutions. He *et al*. (2005) also present a new method for detecting anomalies by discovering frequent patterns (or frequent item sets) within the data set. The anomalies are defined as the data transactions that contain less frequent patterns in their item sets. The recent surveys on the subject (Chandola *et al*., 2009; Patcha & Park, 2007) note that anomalies detection has traditionally dealt with record or transaction type data sets. They further indicate that most techniques require the entire test data before detecting anomalies, and mention very few online techniques. Indeed, most current algorithms assume that the data set fits in the main memory (Yankov *et al*., 2007). Both aspects violate the requirement for real-time monitoring data streams. In addition, most approaches focus specifically on intrusion detection (Kuang & Zulkernine, 2008; Xu *et al*., 2005; Lee & Stolfo, 2000). A comparative study (Chandola *et al.*, 2008) of methods for detecting anomalies in symbolic data shows that there are several techniques for obtaining a symbolic representation from a time series (Lin *et al*., 2007; Bhattacharryya & Borah, 2004), but all such works seem to apply solely to univariate data (Keogh *et al*., 2004; Wei *et al*., 2003). It is a challenging task to detect failures in large dynamic systems because anomalous events may appear rarely and do not have fixed signatures.

9. ANOMALIES DETECTION AND LINK MINING

The literature review reveals a growing range of applications in anomalies detection, mostly to data mining and very few applications in link mining. In recent years application of anomalies detection in link mining has gained increasing importance. For example, the paper of Savage *et a*l (2014) in online social networks survey's existing computational techniques used to detect irregular or illegal behaviour; other works include detecting fraudulent behaviour of online auctioneers (Chan *et al.,* 2006). Community based anomalies detection in evolutionary networks (Chen *et al*., 2012), link based approach for bibliometric journal ranking (Su *et al.,* 2013). However, their focus is still on pattern finding rather than link related tasks. Even the work on citation data (Keane, 2014, Yang *et al*., 2011) is used to describe communities or computational techniques and not mining anomalies or predictive links. Thus, much of the work in this area has focused on identifying patterns in behaviour of the data rather than link mining. Anomalies detection in link mining is still an emerging area.

10. SUMMARY

Link mining is an emerging area within knowledge discovery focused on mining task relationship by exploiting and explicitly modelling the links among the entities. We have overviewed link mining in terms of object related task, link-based object and group related task. These represent some of the common threads emerging from 9 varieties of fields that are exploring this exciting and rapidly expanding field. However, with the introduction of links, new tasks also come to light: predicting the type of link between two objects, predicting the numbers of links, inferring the existence of a link, and inferring the identity of an object. A review of computational techniques is provided outlining their challenges. Anomaly detection, which is important to use in this research, is also discussed and the current methods and issues highlighted. These two areas are attracting much interest by researchers from different disciplines (*e.g.* computer science, business, statistics, forensics and social sciences) interested in extracting tacit, hidden, but valuable knowledge from the vast amount of data available worldwide. Many real-world applications produce data which links to other data, such as the World Wide Web (hypertext documents connected through hyperlinks), social networks (such as people connected by friendship links) and bibliographic networks (nodes corresponding to authors, papers and the edges corresponding to cited-by).

REFERENCES

[1] Abe S., Kawano H., Goldstein J., Ohtani S., Solovyev S.I., Baishev D.G. and Yumoto K. (2006) Simultaneous identification of a plasmaspheric plume by a ground magnetometer pair and IMAGE Extreme Ultraviolet Imager. Journal of Geophysical Research 111(A11).

[2] Aggarwal C., and Yu P.(2001) Outlier Detection for High Dimensional Data. International Conference on Management of Data. 30(2). P.37 – 46.

[3] Aggarwal R, Isil E, Miguel A. Ferreira, and Matos P.(2011) Does Governance Travel Around the World? Evidence from Institutional Investors, Journal of Financial Economics 100. P.154-181.

[4] Aggarwal R., Gehrke J., Gunopulos D.,and Raghavan P.(1998) Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. Proceedings of the ACM SIGMOD International Conference on Management of Data. 27(2). p.94 – 105.

[5] Aggarwal Y. Zhao, and Yu P.S.(2011) Outlier Detection in Graph Streams, ICDE Conference.

[6] Allan J., Carbonell J., Doddington G., Yamron J., and Yang Y. (1998) Topic detection and tracking pilot study. Proceedings of DARPA Broadcast News Transcription and Understanding Workshop.

[7] Badia A., Kantardzic M.(2005) Link Analysis Tools for Intelligence and Counterterrorism. ISI. P.49-59.
[8] Barbara D., Li Y., Couto J., Lin J. L., and Jajodia S.(2003) Bootstrapping a data mining intrusion detection system. Proceedings of the 2003 ACM symposium on Applied computing. ACM Press.
[9] Bhattacharyya N, Bandyopadhyay R, Bhuyan M, et al (2005) correlation of multi-sensor array data with taster's panel evaluation. Proceedings of ISOEN, Barcelona, Spain.
[10] Brockett P. L., Xia, X., and Derrig R. A.(1998) Using kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. Journal of Risk and Insurance. 65(2) P.245-274.
[11] Chandola V., Banerjee A., and Kumar V.(2009) Anomaly Detection. A Survey, ACM Computing Survey. 41(3). p.15.
[12] Chandola V., Eilertson E., Ertoz L., Simon,G., and Kumar V.(2006) Data mining for cyber security. Data Warehousing and Data Mining Techniques for Computer Security, A. Singhal, Ed. Springer.
[13] Chau D. H., Pandit S., Faloutsos C.(2006) Detecting fraudulent 1032 personalities in networks of online auctioneers. In: Knowledge Discovery in Databases: PKDD.p.103–114.
[14] Chellappa., Rama J., and Anil.(1993) Boston: Academic Press.
[15] Chen Z., Hendrix W., Samatova N. F.(2012) Community-based anomaly detection in evolutionary networks. Journal of Intelligent Information Systems. 39(1).p.59–85.
[16] Cox C., Enno A., Deveridge S., Seldon M., Richards R., Martens V., and Woodford P.(1997) Remote electronic blood release system. Transfusion.37.p.960-974.
[17] Creamer, G., and Stolfo, S.(2009) A link mining algorithm for earnings forecast and trading Data. Min Knowl Disc. 18. P.419–445.
[18] Desforges D. M., Lord C. G., Ramsey S. L.(1998) Effects of structured cooperative contact on changing negative attitudes toward stigmatized social groups. Journal of Personality and Social Psychology.60.p.531 -544.
[19] DesJardins M., and Matthew E.(2006) Gaston, Speaking of relations: Connecting statistical relational learning and multi-agent systems. ICML Workshop on Open Problems in Statistical Relational Learning, Pittsburgh, PA.
[20] Domingos P., Doan AH., Madhavan J., and Halevy A.(2004) Ontology matching: A machine learning approach. Handbook on ontologies.
[21] Duda R. O., Hart P. E., and. Stork D. G. (2000) Pattern Classification and Scene Analysis, John Wiley & sons.
[22] Emamian V., Kaveh M., and Tewfik A.(2000) Robust clustering of acoustic emission signals using the kohonen network. Proceedings of the IEEE International Conference of Acoustics,Speech and Signal Processing. IEEE Computer Society.
[23] ErtÄoz A., Arnold M., Prerau L., Portnoy., and Stolfo S.(2003) A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In Proceedings of the Data Mining for Security Applications Workshop.
[24] Ertoz L.; Steinbach, M.; Kumar V.(2004). Finding Topics in collections of documents: A shared nearest neighbour approach. Clustering and Information Retrieval. P.83-104.
[25] Eskin, E., Arnold, A., Prerau, M., Portnoy, L., and Stolfo, S.(2002) A geometric framework for unsupervised anomaly detection. Proceedings of Applications of Data Mining in Computer Security. Kluwer Academics.P.78-100.
[26] Ester M., Kriegel H-P., Sander J., and Xu X.(1996) A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. P.226 – 231.
[27] Fawcett T., Provost F.(1999) Activity monitoring: noticing interesting changes in behavior. Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining (KDD-99).p.53–62.
[28] Fujimaki R., Yairi T., and Machida K.(2005) An approach to spacecraft anomaly detection problem using kernel feature space. Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM Press, New York.p.401–410.
[29] Getoor L.(2003). Link mining. A new data mining challenge, SIGKDD Explorations, 5(1). p.84-89.
[30] Getoor L.(2005) .Tutorial on Statistical Relational Learning. ILP: 415.

[31] Getoor L., and Diehl C.(2005). Link mining: A survey SIGKDD Explorations, December. Vol.7 (2).
[32] Ghosh S., and Reilly D. L.(1994). Credit card fraud detection with a neural-network. Proceeding of the 27th Annual Hawaii International Conference on System Science.3.
[33] Goldberger, a.l.,amaral,A.N.,Glass,L.,Havs dorff,J.M.,ivanov,pc.,mark,R.G.,et al.(2000) physiobank, physiotoolkit and phyionet.circulation,101, 215-220.
[34] Grubbs Frank E.(1969) Procedures the data to assure that the results for detecting outlying observations in are representative of the thing samples. Technometrics 11.p.1-2.
[35] Guha S., Rastogi R., and Shim K.(2001). ROCK: A robust clustering algorithm for categorical attributes. Information Systems 25(5). p. 345-366.
[36] Han J., and Kamber M.(2001) Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers. P.550.
[37] Harris T. (1993). Neural network in machine health monitoring. Professional Engineering.
[38] Harvey Motulsky (1995). Intuitive Biostatistics. Newyork: Oxford University Press. 200-386.
[39] Hu T., and Sung S.Y.(2003) Detecting pattern-based outliers. Pattern Recognition Letters.24 (16). P.3059 – 3068.
[40] Jin W., Tung A., and Han J.(2001). Mining Top-n Local Outliers in Large Databases. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.p.293 – 298.
[41] Keane M., (2014). (Big) Data Analytics: From Word Counts to Population Opinions. insight. 1,1-45.
[42] Keogh E., Lonardi S., and Ratanamahatana C. A.(2004). Towards parameter-free data mining. Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press.p. 206-215.
[43] Kirkland, D., Senator, T., Hayden, J., Dybala, T., Goldberg, H. & Shyr, P. (1999). The NASD Regulation Advanced Detection System. AAAI 20(1): Spring, 55-67.
[44] Kohonen T.(1997) Self-organizing maps. Springer-Verlag New York, Inc.
[45] Kuang L., and Zulkernine M.(2008) An anomaly intrusion detection method using the CSI-KNN algorithm. SAC.P. 921-926.
[46] Labib K., and RaoVemuri V.(2002) "NSOM: A Real-time Network-Based Intrusion detection System Using Self-Organizing Maps, Networks and Security.
[47] Lakhina A., Crovella M., and Diot C.(2005) Mining Anomalies Using Traffic Feature Distributions. Proceedings of ACM SIGCOM.p. 217-228.
[48] Lee W., and Stolfo, S.(2000) A framework for constructing features and 638 models for intrusion detection systems. ACM Transactions on 639 Information and System Security. 3(4).
[49] Liben-Nowell D., and Kleinberg J.(2003) The link prediction problem for social networks. In CIKM '03. Proceedings of the twelfth international conference on Information and knowledge management.p.556–559.
[50] Lin H., Fan W., and Wallace L.(2007) An empirical study of web-based knowledge community success. Proceedings of the 40th Hawaii International Conference on System Sciences. P.1530-160.
[51] Lin S., and Brown D.(2004) An Outlier-based Data Association Method for Linking Criminal Incidents. Proceedings of the SIAM International Conference on Data Mining.
[52] Lin S., and Brown D.(2003) An Outlier-based Data Association Method. Proceedings of the SIAM International Conference on Data Mining.
[53] Marchette D.(1999) A statistical method for profling network traffic. Proceedings of 1st USENIX Workshop on Intrusion Detection and Network Monitoring. Santa Clara, CA, .p.119-128.
[54] Marcus G., Fernandes K., and Johnson S.(2007) Infant rule-learning facilitated by speech. Psychol. Sci.18.p.387–391.
[55] Mustafa Y. T., Tolpekin V., and Stein A., and Sub M.(2007) The application of Expectation Maximization algorithm to estimate missing values in Gaussian Bayesian network modeling for forest growth. IEEE Transactions on Geoscience and Remote Sensing.
[56] O'Madadhain J., Smyth P., and Adamic L.(2005) Learning Predictive Models for Link Formation. To be presented at the International Sunbelt Social Network Conference.
[57] Otey M. E., Ghoting A., and Parthasarathy S.(2003) Fast distributed outlier detection in mixed-attribute data sets. Data Mining and Knowledge Discovery. 12(2-3) p.203-228.

[58] Otey M., Parthasarathy S., Ghoting A., Li G., Narravula S., and Panda D.(2003).
[59] P´oczos Z., and L˝orincz A.(2009) Complex independent process analysis. PLA University of Science & Technology, Nanjing 210007, China.
[60] Panzeri S., Brunel N., Logothetis NK., and Kayser C.(2010) Sensory neural codes using multiplexed temporal scales. Trends Neurosci.33.p.111–120.
[61] Patcha A., and Park JM.(2007) An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer Networks. 51(12).p.3448-3470.
[62] Petrovskiy M.(2003) Outlier Detection Algorithms in Data Mining Systems. Programming and Computing Software. 29(4).p.228 – 237.
[63] Pires A., and Santos-Pereira C.(2005) Using clustering and robust estimators to detect outliers in multivariate data. Proceedings of International Conference on Robust Statistics. Finland.
[64] Platt J.(2000) Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. A. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans, Eds.p.61–74.
[65] Popescul A., Ungar L., Lawrence S., and Pennock D.(2003) Statistical re-lational learning for document mining. Computer and Information Sciences, University of Pennsylvania.
[66] Provana K.G., Leischowc S. J., Keagyb J., and Nodorac J.(2010) Research collaboration in the discovery, development, and delivery networks. of a statewide cancer coalition.33(4).p. 349-355.
[67] Ramadas M., Ostermann S., and Tjaden B. C.(2003) Detecting anomalous network traffic with self-organizing maps. Proceedings of Recent Advances in Intrusion Detection.P.36-54.
[68] Ramaswamy S., Rastogi R., and Shim K.(2000). Efficient Algorithms for Mining Outliers from Large Data Sets. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. 29(2).p.427 – 438.
[69] Rattigan M. J., and Jensen D.(2005) The case for anomalous link discovery. SIGKDD Explorations, 7(2).
[70] Savage D., Zhang X., Yu X., Chou P., and Wang Q.(2014) Anomaly Detection in Online Social Networks. Social Networks.39.p.62–70.
[71] Scarth G., McIntyre M., Wowk B., and Somorjai R.(1995) Detection of novelty in functional images using fuzzy clustering. Proceedings of the 3rd Meeting of International Society for Magnetic Resonance in Medicine. Nice, France.p.238.
[72] Sequeira K. and Zaki M.(2002) Admit: anomaly-based data mining for intrusions. In Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press.p.386–395.
[73] Sheikholeslami G., Chatterjee S., and Zhang A.(1998) Wavecluster: A multi-resolution clustering approach for very large spatial databases. Proceedings of the 24rd International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc.p.428-439.
[74] Shyu M.L., Chen S.C., Sarinnapakorn K., and Chang L.(2003) A novel anomaly detection scheme based on principal component classifier. Proceedings of 3rd IEEE International Conference on Data Mining.p.353–365.
[75] Skillicorn D. B.(2004) Detecting Related Message Traffic, Workshop on Link Analysis, Count ErtÄoz errorism, and Privacy. SIAM International Conference on Data Mining, Seattle, USA.
[76] Smith R., Bivens A., Embrechts M., Palagiri C., and Szymanski B.(2002) Clustering approaches for anomaly based intrusion detection. Proceedings of Intelligent Engineering Systems through Artificial Neural Networks. ASME Press.P.579-584.
[77] Song X., Wu M., Jermaine C., and Ranka S.(2007) Conditional anomaly detection. IEEE Transactions on Knowledge and Data Engineering.19(5).p.631-645.
[78] Sparrow M.(1991) The application of network analysis to criminal intelligence: an assessment of the prospects. Soc Netw 13.p.251–274.
[79] Su, P., Shang C., and Shen A.(2013) Soft Computing - A Fusion of Foundations, Methodologies and Applications archive. 17(12).p.2399-2410.
[80] Su, P., Shang C., and Shen A.(2013) "Link-based approach for bibliometric journal ranking," Soft Computing, to appear.
[81] Tan L.,Taniar D., and Smith K.(2005) Introduction to Data Mining. Addison-Wesley.J(2).p.229-245.

[82] Taskar B., Abbeel P., and Koller D.(2003) Discriminative probabilistic models for relational data. Proc. UAI02, Edmonton, Canada.
[83] Thottan., and Ji.(2003) Anomaly detection in IP networks. Signal Processing, IEEE Transactions .51(8).p.2191-2204.
[84] Vinueza A., and Grudic G.(2004) Unsupervised outlier detection and semi-supervised learning.Tech. Rep. CU-CS-976-04, Univ. of Colorado at Boulder.
[85] Wadhah,A.,Gao,S., Jarada,T., Elsheikh,A.,Murshed,A.,Jida,J.,Alhajj,R.. (2012). Link prediction and classification in social networksand its application in healthcare and systems biology. Netw Model Anal Health Inform Bioinforma. 1 (2), 27-36.
[86] Wei L., Qian W., Zhou A., and Jin W.(2003). Hot: Hypergraph-based outlier test for categori-cal data. Proceedings of the 7th Pacic-Asia Conference on Knowledge and Data Discovery. p.399-410.
[87] Wu N. , and Zhang J.(2003) Factor analysis based anomaly detection. Proceedings of IEEE Workshop on Information Assurance. United States Military Academy, West Point, NY.
[88] Wu J., Xiong H., and Chen J.(2009) .Adapting the right measures for k-means clustering, in KDD.p.877–886.
[89] Xu, K.M., Zhang M, Eitzen Z.A., Ghan S.J., Klein S.A., and Zhang J.(2005) Modeling springtime shallow frontal clouds with cloud-resolving and single-column models. J. Geophys. Res., 110, D15S04, doi:10.1029/2004JD005153.
[90] Yankov D., Keogh E. J., and Rebbapragada U. (2007). Disk aware discord discovery: Finding unusual time series in terabyte sized datasets. Proceedings of International Conference on Data Mining.p.381-390.
[91] Yang Y., Zhiguo G., and Leong H.U.(2011). Identifying points of interest by self-tuning clustering. Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR '11). ACM, New York.
[92] Ypma A., and Duin R.(1998). Novelty detection using self-organizing maps. Progress in Connectionist Based Information Systems.2.p.1322-1325.
[93] Yu D., Sheikholeslami G., and Zhang A.(2002) Findout: finding outliers in very large datasets. Knowledge And Information Systems.4(4).p. 387.