

LINK MINING PROCESS

Dr.Zakea Il-Agure and Mr.Hicham Noureddine Itani

Higher Colleges of Technology, United Arab Emirates

ABSTRACT

Many data mining and knowledge discovery methodologies and process models have been developed, with varying degrees of success, there are three main methods used to discover patterns in data; KDD, SEMMA and CRISP-DM. They are presented in many of the publications of the area and are used in practice. To our knowledge, there is no clear methodology developed to support link mining. However, there is a well known methodology in knowledge discovery in databases, known as Cross Industry Standard Process for Data Mining (CRISPDM), developed by a consortium of several industrial companies which can be relevant to the study of link mining. In this study CRISP-DM has been adapted to the field of Link mining to detect anomalies. An important goal in link mining is the task of inferring links that are not yet known in a given network. This approach is implemented through the use of a case study of realworld data (co-citation data). This case study aims to use mutual information to interpret the semantics of anomalies identified in co-citation, dataset that can provide valuable insights in determining the nature of a given link and potentially identifying important future link relationships.

KEYWORDS

Link mining, anomalies, mutual information

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.

In the last decade we have seen an increasing interest in the study of anomalies detection in data mining applied to law enforcement, financial fraud, and terrorism. In recent years, this study has been applied to social networks and online communities to identify influential networks participants and predict fraudulent or malicious activities.

To our knowledge, the study of anomaly detection in link mining relied mostly on statistical or machine learning methods in order to gain insight to the structure of their networks. We believe that we can achieve a better understanding of these anomalies if we apply mutual information to the data entities and objects and links to reveal their semantic relationship. The aim of this research is to show how mutual information can help in providing a semantic interpretation of anomalies in data, to characterise the anomalies, and how mutual information can help measuring the information that object item X shares with another object item Y. This paper attempted to demonstrate the contribution of mutual information to interpret anomalies using a case study. This paper presents a novel approach to anomaly detection in link mining methodology based on mutual information.

2. LINK MINING METHODOLOGY

As CRISP-DM methodology is well developed and applied in knowledge discovery; this research has adapted it to the emerging field of link mining. While data mining addresses the discovery of patterns in data entities, link mining is interested in finding patterns in objects by exploiting and modeling the link among the objects. The approach to link mining is still an ad-hoc approach. The proposed adopted CRISP-DM methodology can help provide a structured approach to link mining in Figure 1. This consists of six stages:

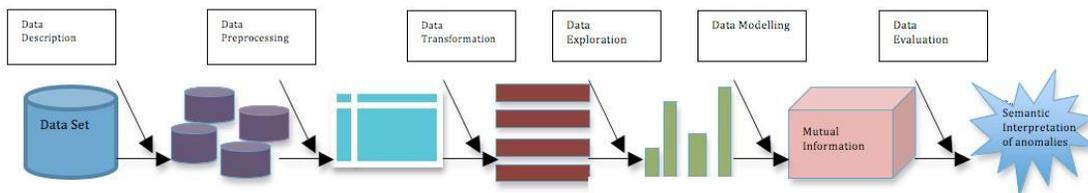


Figure 1. Link mining methodology

The aim of this methodology is to define the link mining task and determining the objectives of link mining.

1. Data description. The data description phase starts with initial data collection and proceeds with activities that enable the researcher to become familiar with the data. The aim is to check data quality and any associated problems in order to discover first insights into the data, and identify interesting subsets to form hypotheses regarding hidden information.
2. Data pre-processing. The data pre-processing phase covers activities related to data cleansing and data integrity needed to construct the final dataset from the initial raw data. While outliers can be considered noise, or anomalies and thus discarded in data mining, they become the focus of this study as they can reveal important knowledge in link mining.

3. Data transformation. This involves syntactic modifications applied to the data; this may be required by the modelling tool. Selecting an appropriate representation is an important challenge in link mining. The objects in link mining (e.g. people, events, organisation, and countries) have to be transformed into feature factors to represent and capture the connectivity and the strength of the links among those objects.
4. Data exploration. This stage is concerned with the distribution of the data and using relevant graphical tools to visualise the structure of the objects and their links. This stage helps identify the existence of anomalous objects or links.
5. Data modelling. This stage aims to identify all entities and the relationship between them. Data modeling puts algorithm in general in a historical perspective rooted in mathematics, statistics, and numerical analysis. For more complex data sets, different techniques are used such as nearest neighbour, statistical, classification, and information/context based approaches.
6. Evaluation: Data cleaning solutions will clean data by cross checking with a validated data set in phase 2. The clustering model in phase 5, explains natural groupings within a dataset based on a set of input variables. The resulting clustering model is sufficient statistics for calculating the cluster group norms and anomaly indices. Mutual information is useful in validating the model as it provides a semantic underpinning to the patterns and discoveries made in phase 5.

3. CASE STUDY

The application of the novel approach is implemented to a case study to demonstrate how mutual information can help explore and interpret anomalies detection with a real-world data set and application area. The key challenge for this technique is to apply data representation, for example graphs to visualise the dataset and a clustering approach (hierarchical cluster method). In Figure 2 shows how this study focuses on a case study using a set of co-citation data. The link mining methodology described above is applied to this case study and includes the following stages: data description, data preprocessing, data transformation, data exploration, data modeling based on graph mapping, hierarchical cluster and visualisation, and data evaluation.

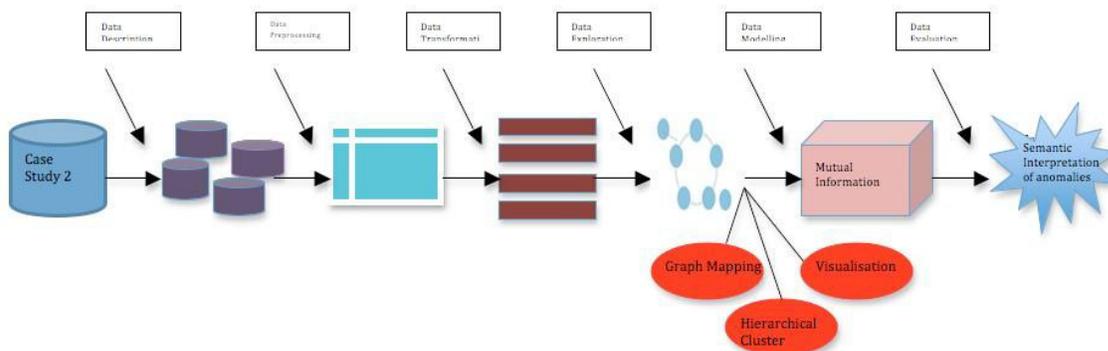


Figure 2. Link mining methodology in case study

This case study covers the three link mining tasks. It is an attempt at identifying and clustering objects, representing them into a graph structure and studying the links between these objects.

4. DISCUSSION

If the approach were to be valid when used with a data set where the anomalies and relationships are unknown, it is necessary to demonstrate that the approach could be scaled to real world data volumes and used with inconsistent and/or noisy data and with other clustering algorithms. This case study addresses these issues. The clustering approach used in this case study was hierarchical clustering. Using bibliographic data, this approach created 5 clusters. Cluster 1 was found to contain data with the strongest links and cluster 5 to contain data with the weakest links. Applying mutual information, we were able to demonstrate that the clusters created by applying the algorithm reflected the semantics of the data. Cluster 5 contained the data with the lowest mutual information calculation value. This demonstrated that mutual information could be used to validate the results of the clustering algorithm.

As the result in Table 1, cluster 1 shows high mutual information indicating higher co-citation strength; cluster 5 has low mutual information indicating lower co-citation strength.

Table 1. Result of mutual information

	Clusters	Items	Colour	Mutual information
1	Cluster1	58		0.93
2	Cluster2	49		0.82
3	Cluster3	38		0.63
4	Cluster4	29		0.43
5	Cluster5	19		0.00

It was necessary to establish whether the proposed approach would be valid if used with a data set where the anomalies and relationships were unknown. Having clustered and then visualized the data and examined the resulting visualisation graph and the underlying cluster through mutual information, we were able to determine that the results produced were valid, demonstrating that the approach can be used with the real world data set. Analyzing each of the clusters, and the relationships between elements in the clusters was time consuming but enabled us to establish that the approach could be scaled to real world data and that it could be used with anomalies which were previously unknown. We found with the case study that the semantic preprocessing stage was an essential first step. The data from the bibliographic sources normally contains errors, such as misspelling the author's name, the journal title, or in the references list. Occasionally, additional information has to be added to the original data, for example, if the author's address is incomplete or wrong. For this reason, the analysis cannot be applied directly to the data retrieved from the bibliographic sources -a pre-processing stage over the retrieved data is necessary to overcome these issues. In this case study, the clustering approach was used to cluster the data into groups sharing common characteristics, graph based visualization and mutual information was used to validate the approach.

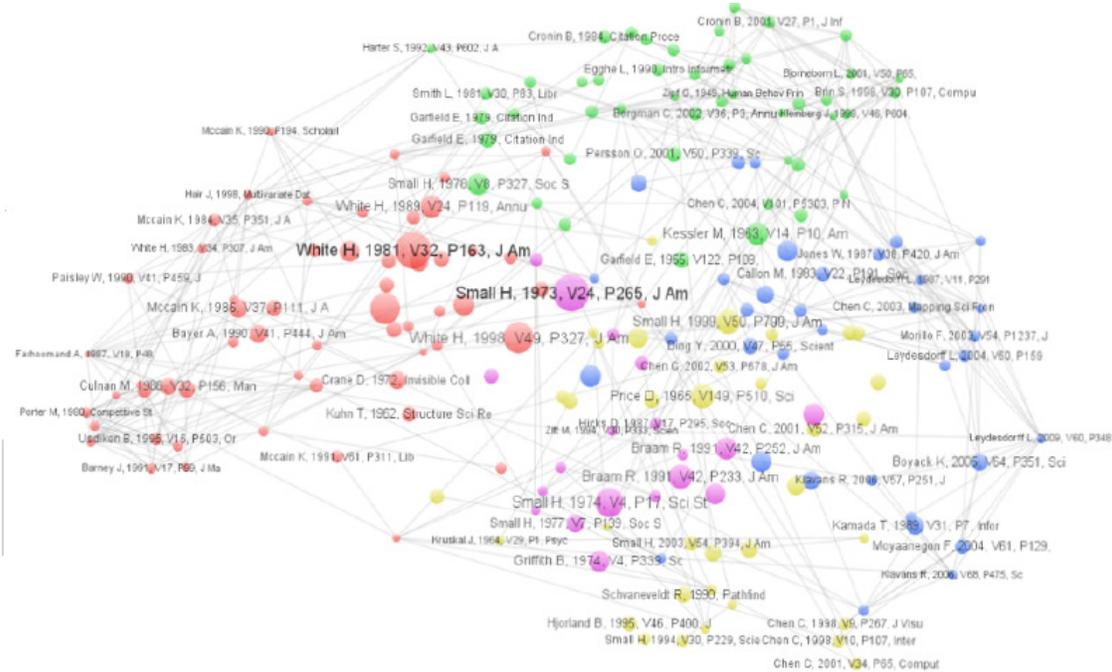


Figure 3. Mapping nodes

Clusters are designed to classify observations, as anomalies should fall in regions of the data space where there is a small density of normal observations. The anomalies occur in this case study as a cluster among the data, such observations are called collective anomalies, defined by Chandola et al. (2009) as follows: “The individual data instances in a collective anomaly may not be anomalies by themselves, but their occurrence together, as a collection is anomalous.” Existing work on collective anomaly detection requires supporting relationships to connect the observations, such as sequential data, spatial data and graph data. Mutual information can be used to interpret collective anomalies. Mutual information can contribute to our understanding of anomalous features and help to identify links with anomalous behaviour. In this case study, mutual information was applied to interpret the semantics of the clusters. In cluster 5, for example, mutual information found no links amongst this group of nodes. This indicates collective anomalies, as zero mutual information between two random variables means that the variables are independent. Link mining considers data sets as a linked collection of interrelated objects and therefore it focuses on discovering explicit links between objects. Using mutual information allows us to work with objects without these explicit links. Cluster 5 contained documents, which had been selected as part of the co-citation data, but these documents were not themselves cited. Mutual information allowed us to examine the relationships between documents and to determine that some objects made use of self-citation meaning that they were regarded co-cited but did not connect to other objects. We also identified a community anomaly, where the edge is considered a relationship anomaly, because it connects two communities, which are usually not connected to one another. Mutual information provided information about the relationships between objects, which could not be inferred from a clustering approach alone. This additional information supports a semantic explanation of anomalies.

The co-citation data applied hierarchical clustering and visualized the data as a graph where nodes represented authors and edges represented cited by. The aim was to cluster the nodes into

groups sharing common characteristics; mutual information was applied to all clusters and demonstrated strong links among the element of each cluster, except in cluster 5. Mutual information conforms that cluster 5 elements share no links with the clusters and among themselves no link was found between authors. Zero mutual information between two random variables means that the variables are independent. As the discussion shows, mutual information can provide a semantic interpretation of anomalous features.

5. SUMMARY

In this study, hierarchical clustering is applied to identify clusters and the data is visualised using graph representation. Anomalies occur as a cluster among the data, such observations are collective anomalies. Cluster validity with respect to anomalies can be difficult to evaluate because of data volumes. This research has demonstrated that mutual information can be applied to evaluate cluster content and the validity of the clustering approach. This also supports validation of the visualisation element. This case study was developed to use mutual information to validate the visualization graph. We used a real world data set where the anomalies were not known in advance and the data required pre-processing. We were able to show that the approach developed when scaled to large data volumes and combined with semantic pre-processing, allowed us to work with noisy and inconsistent data. The co-citation data applied hierarchical clustering and visualised the data as a graph where nodes represented authors and edges represented cited-by. The aim was to cluster the nodes into groups sharing common characteristics; mutual information was applied to all clusters and demonstrated strong links among the element of each cluster, except in cluster 5. Mutual information conforms that cluster 5 elements share no links with the clusters and among themselves no link was found between authors. Zero mutual information between two random variables means that the variables are independent. Mutual information supported a semantic interpretation of the clusters, as shown by the discussion of cluster 5. The experimental work confirmed the effectiveness and efficiency of the proposed methods in practice.

In particular, this revealed that our method is able to deal with data sets with a large number of objects and attributes. Having clustered and then visualised the data and examined the resulting visualisation graph and the underlying cluster through mutual information, we were able to determine that the results produced were valid, demonstrating that the approach can be used with the real world data set. Anomalies detection finds applications in many domains, where it is desirable to determine interesting and unusual events in the activity, which generates such data. The core of all anomalies detection methods is the creation of a probabilistic, statistical or algorithmic model, which characterises the normal behavior of the data. The deviations from this model are used to determine the anomalies. A good domain-specific knowledge of the underlying data is often crucial in order to design simple and accurate models, which do not over fit the underlying data. Using mutual information contributes to our understanding of the anomalous features and helps with semantic interpretation and to identify links with anomalous behavior. The problem of anomalies detection becomes especially challenging, when significant relationships exist among the different data points. This is the case for bibliographic data in which the patterns in the relationships among the data points play a key role in defining the anomalies. In the data used in this case study, there is significantly more complexity in terms of how anomalies may be defined or modelled which can be used to interpret semantic meaning. Therefore, anomalies may be defined in terms of significant changes in the underlying network community or distance structure. Such models combine network analysis and change detection in

order to detect structural and temporal anomalies from the underlying data. This research has demonstrated that mutual information can be applied to evaluate cluster content and the validity of the clustering approach. This also supports validation of the visualization element.

REFERENCES

- [1] G. Chandola V., Banerjee A., and Kumar V.(2009) Anomaly Detection. A Survey, ACM. Computing Survey. 41(3). p.15.
- [2] Shearer C., The CRISP-DM model: the new blueprint for data mining, J Data Warehousing (2000); 5:13—22..
- [3] IL-agure, Z. I. (2016). Anomalies in link mining based on mutual information). Staffordshire University. UK.