INTERPRETING THE SEMANTICS OF ANOMALIES BASED ON MUTUAL INFORMATION IN LINK-MINING

Dr. Zakea Il-agure and Dr. Belsam Attallah

Department of Computer Information Science, Higher Colleges of Technology, United Arabs Emirates

ABSTRACT

This paper aims to show how mutual information can help provide a semantic interpretation of anomalies in data, characterize the anomalies, and how mutual information can help measure the information that object item X shares with another object item Y. Whilst most link mining approaches focus on predicting link type, link based object classification or object identification, this research focused on using link mining to detect anomalies and discovering links/objects among anomalies. This paper attempts to demonstrate the contribution of mutual information to interpret anomalies using a case study.

KEYWORDS

Anomalies, Mutual information, Link mining, co-citation

1. INTRODUCTION

Link mining refers to data mining techniques that explicitly consider links when building predictive or descriptive models of linked data. Getoor and Diehl (2005) identify a set of commonly addressed link mining tasks, which are: *Object-related tasks, Link-related tasks* and *Graph-related tasks* (which has been used in this case study).

Link mining is a newly developed research area, bringing together research insights from the fields of web mining, graph theory and machine learning. Link mining applications have been shown to be highly effective in addressing many important business issues such as telephone fraud detection (Fawcett & Provost, 1999), crime detection (Sparrow, 1991), money laundering (Kirkland et al., 1999), terrorism (Badia & Kantardzic, 2005; Skillicorn, 2004), financial applications (Creamer & Stolfo, 2009), social networks and health care problems (Provana et al., 2010; Wadhah et al., 2011). The trend in the building and use of link mining models for critical business, law enforcement and scientific decision support applications are expected to grow. An important issue will be building models and techniques that are scalable and reliable. 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

23

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). This paper aims to use mutual information to interpret the semantics of anomalies identified in co-citation dataset which can provide valuable insights in determining the nature of a given link and potentially identifying important future link relationships. The case study is used to demonstrate how mutual information can help explore and interpret anomalies detection using a set of co-citation data. The key challenge for this technique is to apply the approach to real world data set, making use of a different form of data representation, for example graphs to visualise the dataset. The link mining methodology described (IL-agure, 2016) is applied to the case study and includes the following stages: data description, data pre-processing, data transformation, data exploration, data modelling based on graph mapping, hierarchical cluster and visualisation, and data evaluation
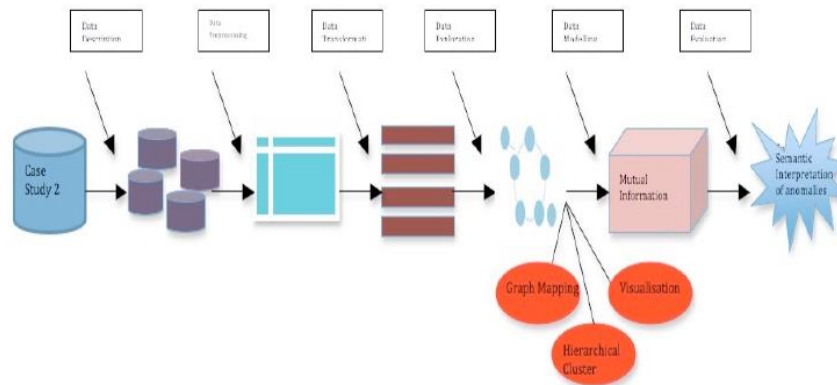.

Figure 1. Link mining methodology

The Stage1: Data description

There are several online bibliographic databases where scientific works, documents and their citations are stored. The most important bibliographic databases are the *Web of Science* ISI (WoS), *Scopus*, and *Google Scholar*. This case study extracted 569 records, from *Web of Science*, and stored them in a spreadsheet file. These 569 records include 1001 co-citations from three databases: SCI-EXPANDED, SSCI, A&HCI up to 2011. Each co-citation include the author 'name, journal, cited documents and cited references. The author is the entity that signifies the person who has been involved in the development of the document. An author can be linked to a set of documents, and in a similar way, a document has a group of authors. Also, an author has a linked position in his/her documents. Pairs of citations being cited by a common citing document identified co-citation relationships. The strength of the relationship is based on the number of citing documents that contain the citations. The chance of citations being co-cited increases based on the number of times the citation appears in reference lists of citing documents. Citations contained in a large number of reference lists have a greater chance of being co-cited than citations found in a smaller number of reference lists. Co-citation strength were used to account for the frequencies of citations found in the reference lists of citing documents.

Stage 2: Data pre-processing

The data from the bibliographic sources contain a number of errors, such as misspelling in the author's name, in 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 task over the retrieved data is required, to improve the quality of the data and the analysis. A set of pre-processing tasks is applied to prepare the data and is described

(i)Data reduction: aims to select the most important data, which is normally an extensive task. With such a quantity of data, it could be difficult to obtain good and clear results in the relationship. For this reason, it is often conducted using a portion of the data.

(ii)Detecting duplicate and misspelled items: There are items in the data that represent the same object or concept but with different spelling, for example, an author's name can be written in different ways (e.g., Zakia.Il; Il Agure Zakea), and yet each spelling represents the same author. In other cases, a concept is represented with different words (lexical forms) or acronyms, and yet refers to the same concept. To improve data quality, first authors' initials, are kept and converted from lower to upper case to maintain consistency. The first author 'name is used in our analysis.

Stage 3: Data transformation

Several relations among the nodes can be established. The focus in the case study was on cocitation in the bibliometric technique taxonomy. The similarity between the nodes of analysis is usually measured counting the times that two nodes appear together in the documents. The nodes of analysis used in case study are author, citation document and journal. Different aspects of a research field can be analysed depending on the selected nodes for analysis. Additionally, a link can be used to attain the relation among nodes, the extraction of co-citation network by using *BibExcel*, in order to help with citation studies, and bibliographic analysis, in particular:

1. Convert to dialog format/convert from Web of Science. A bibliographic record consists of a number of fields used to index the actual text, its subjects and descriptive data. When working with BibExcel we usually transform the initial data to the dialog format in Figure 2 more pecifically the format for Science Citation Index. Common data between records are thus structured in univocal metadata fields, such as publication titles in the title field, authors in the author field, and references in the reference filed.

2. Extracting data from CD-field (citation-documents) where the relations of the different entities related with each document (authors, year, vol., page, and journal) are stored.

3. To improve data quality, only the first authors' initials are retained (see Figure 2).
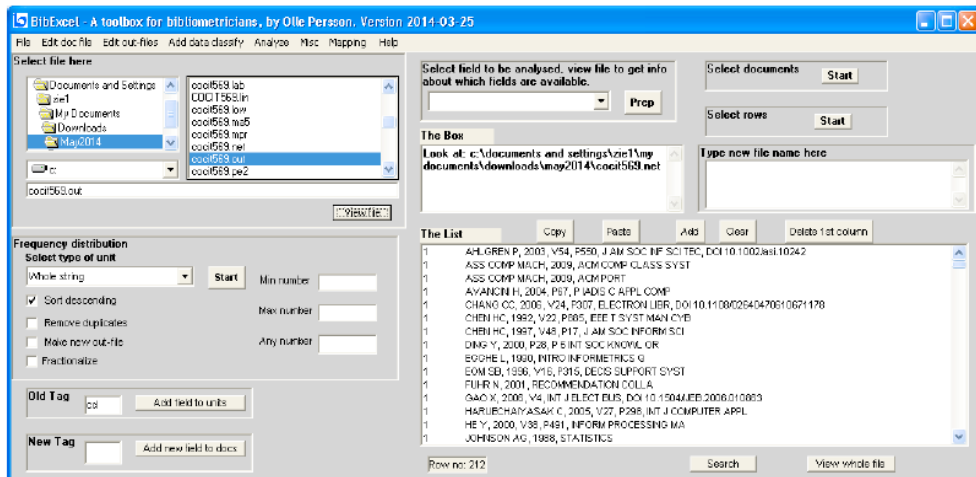
Figure 2. Retaining first authors' initials

Stage 4: Data exploration

Once the network of relationships between the selected nodes has been built, an exploration is applied to the data to derive similarities from the data. For instance, if a co-citation analysis is performed and various clusters are detected, then a label would be set to each one. This label should be selected using the most important document terms of the cluster.

a)Computing frequencies of citations

When making the OUT-file, specific bibliographic fields need to be selected, from which the OUT-file will be constructed. Depending on which bibliographic fields are chosen as a unit when the OUT-file is created, the frequency calculation function in BibExcel offers many different selections. Such as, if the file name: OUT-file consists of a cited document, BibExcel can make a substring search and only count a specified part of the cited document, such as cited author or cited journal.
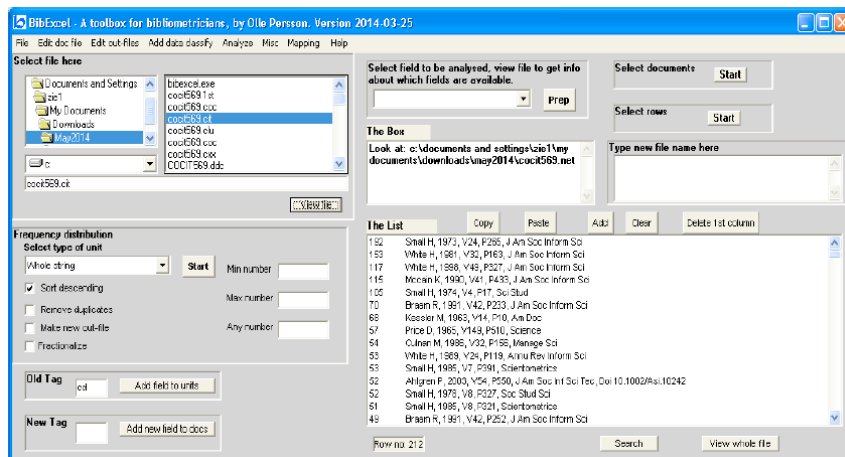

Figure 3. The frequency

b) Making co-citations

Co-citation is a semantic similarity measure for documents that makes use of citation relationships. The definition of co-citation is the frequency with which two documents *are* cited together by other documents (Small, 1973). If at least one other document cites two documents in common these documents are co-cited. The higher the co-citation strength, the more co-citations two documents receive and more likely they are semantically related (see Figure 4).
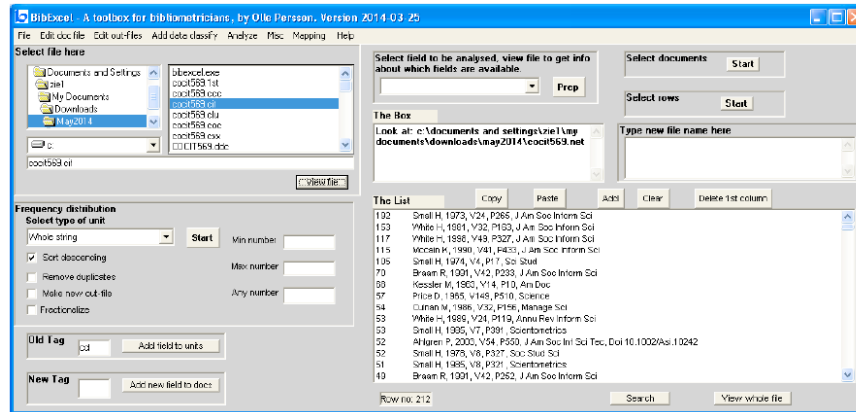

Figure 4. Making co-citations

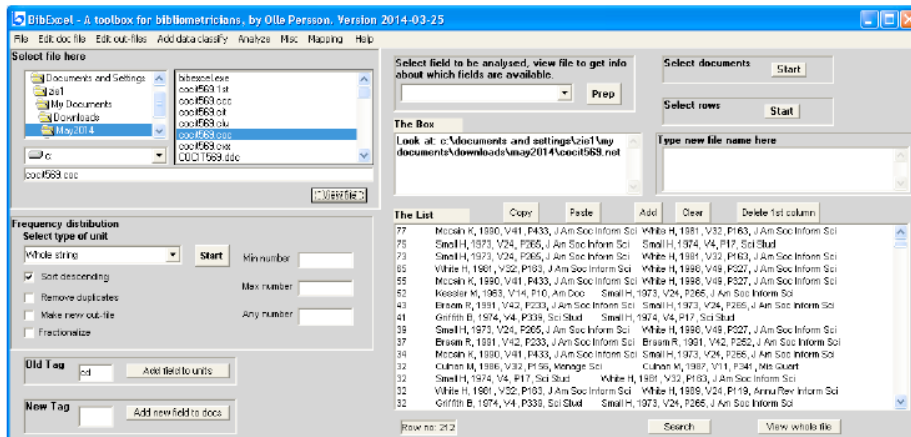3. Make co-occurrences *pairs* via the list box.

Figure 5. Making co-citations pairs

The menu analysis presented contains a number of specialised functions permitting the analyses of citation networks and, perhaps most importantly, a range of different co-occurrence analyses. We will therefore focus on co-occurrence analysis – how to prepare the data and how to perform co-occurrence analyses.

Co-occurrence analysis is the study of mutual appearances of pairs of units over a consecutive number of bibliographic records. Therefore, the unit of analysis in the OUT-file defines the type of co-occurrence analysis. For example, an OUT-file that lists the individual authors from each record in the Doc-file would be the basis for a co-author analysis. The matching routine used to match pairs of units must therefore be performed on the OUT-file. It is the nodes in the individual documents and their frequency across all documents that must be generated.

Stage 5: Data Modeling

The modelling step is the most important stage. The co-cited data is represented first using a graph representation for visualisation purposes. BibExcel is used to produce *net-files* for cocitations, which are converted for further analysis and visualisation with ***VOSviewer*** (See Figure 6). The VOSviewer tool is used to build a map based on a co-occurrence matrix. (Van Eck and Waltman, 2009a, 2009b*).* The VOS viewer map created for case study is given below:
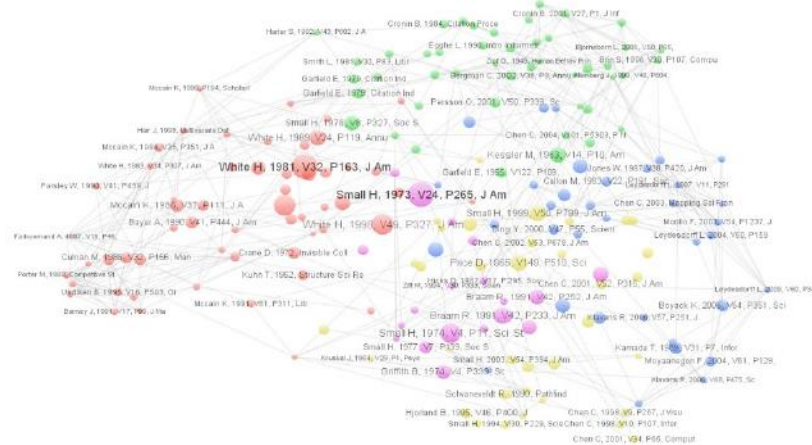

Figure 6. Mapping nodes

1 .Graph analysis of co-citation data

Anomalies represent significant deviations from 'normal' structural patterns in the underlying graphs. This description is lengthy because much is involved in its preparation, measurement, results and expressing the differences between the groups in some way (the statistic test), and choosing an inference procedure built on that statistic. Each pattern is under the control of the experimenter or observer and each is important. The concept of finding a pattern that is 'similar' to frequent, or good pattern is different from most approaches that are looking for unusual or 'bad' patterns. There is no universal definition of the problem, as it depends heavily on: The application domain and the properties in addition to the properties of the graph under consideration.

The main goal of anomalies in graphs is to highlight unusual relationships in the graphs by representing them as edges between regions of the graph that rarely occur together. In a citation network, two co-authors who are drawn from groups that usually do not work together may sometimes publish together (cross-disciplinary papers). Such anomalies provide unique insights about the relationships in the underlying network.

Anomalies may be modelled in different ways depending upon the abnormality of either the nodes in terms of their relationships to other nodes, or the edges themselves. In such cases, in Figure 7 below a node, which illustrates irregularity in its structure within its region, may be considered as an anomaly (Akoglu et al., 2010). Also, an edge which connects different communities of nodes may be considered a relationship or community anomaly (Aggarwal et al., 2011) and (Gao et al., 2010). Figure 7 (a) contains a case of a node anomaly, because node 5 has an unusual locality structure, which is significantly different from the other nodes as (Chen C, 1998, V9, P267, J Visu) in the map. Figure 7 (b) Node 5 is that disconnected and is far away from other cluster members as (Zitt M, 1994, V30, P333, Scien)in the map. On the other hand, the edge (2, 4) in Figure 7 (c) may be considered a relationship anomaly or community anomaly, because it connects two communities, which are usually not connected to one another as (Kessler

M, 19963, V14, P10, Am) in the map. Hence, in the graph data, there is significantly more difficulty and flexibility in terms of how anomalies may be defined or modelled.
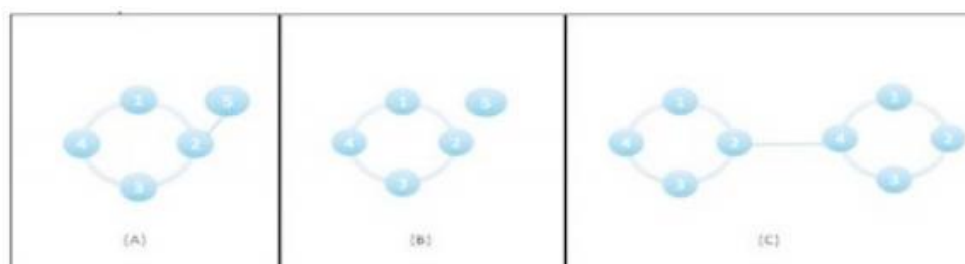

Figure 7. Cases of node anomalies

2. Hierarchical Cluster

A crucial step to evaluate whether mutual information-based measures can be effectively used to represent strength of group ties in network analysis is to examine the extent to which the network structures derived from mutual information-based measures resemble the true network structures. Thus, hierarchical cluster is introduced in the current study for the purpose of network structure inference. Hierarchical cluster is one of the many strategies that have been used to visualise the relationship among elements of a network and to make inference on the overall structure of the network from proximity data among those elements (Aghagolzadeh et al., 2007; DeJordy, et al., 2007; Hubert et al., 2006; Kraskov & Grassberger, 2009; Kraskov et al., 2005). Given aproximity matrix of *n* elements, the primary goal of hierarchical clustering analysis is to find a partition hierarchy. This analysis is usually performed as beginning from a full partition where each element forms a subgroup, elements are grouped together step by step. Clustering algorithms was used to group data into 5 different clusters. The clustering grouped 193 nodes, into 5 clusters. The largest cluster is cluster 1 with 58 items and cluster 5 is the smallest with 19 items.

Co-citation is defined as the frequency with which two documents are cited together by other ocuments. If one other document cites two documents in common these documents are cocited. The higher co-citations two documents receive, the more their co-citation strength, and are semantically related, which can be related to the results from the mapping nodes. Where cluster 1 shows high co-citation frequency indicating higher co-citation strength, cluster 5 has a low cocitation frequency indicating lower co-citation strength. The relationship strength is based on the number of citations the two citing documents have in common. After the creation of author cocitation pairs, the co-citation link strength (Garfield, 1980) is calculated using the following formula：

$$\text{Link Strength}\,(AB) = X / (Y - X)$$

Where X is the number of co-citations of author A and author B, Y is the sum of the total number of citations of A and the total number of citations of B. This formula normalises the co-citation link strength by taking into account the total number of citations for both A and B. In item 1 (Small H, 1973) the link strength is 1818 indicating that it is present in cluster 1 and is more co cited, however item 193 (Farhoomand A, 1987) is shown to have the lowest link strength of 50 and is present in cluster 5 indicating that it less co cited.

3. Visualisation

Analysis of networks has been widely used in a great number of areas to understand relationshipsbetween different entities in a network, as well as behaviour of a network as a whole

due to the interactions between entities within it. Researchers have conducted observations and developed, experiments on a variety of network analysis techniques including graphical visualisation, statistical inference and computational algorithms, and built a number of mathematical models in an effort to understand and predict the behaviour of a network (Newman, 2003). Co-citation data can be used to study relations among authors or journals; it can be used to construct the maps that provide a visual representation of the structure of a scientific field. Usually, when using co-occurrence data, a transformation is applied first to the data. The aim of such a transformation is to derive similarities from the data. For example, when researchers study relations among authors based on co-citation data, they typically derive similarities from the data and then analyse these similarities using hierarchical clustering.

The visualisation helps provide a clear understanding and better representation of the output map represented at co-citation (see Figure 6). The resulting map visualises a set of objects and the relations among the objects. Many different types of visualisations can be used. One difference is between distance-based visualisations and graph-based visualisations. In distance-based visualisations, the distance between two nodes reflects the relation between the nodes. The smaller the distance between two nodes, the stronger the relation between the nodes. On the other hand, in graph-based visualisations in the case study, the distance between two nodes does not reflect the relation of the nodes. Instead, drawing lines between nodes from the map typically indicates relations between nodes; the most basic way to visually group nodes is to use colours. If items have been assigned to clusters, the colour of the circle of an item can be determined by the cluster in which the item belongs. Item cluster is calculated and translated into colours using a colour scheme. By default, VOSviewer uses a red-green- blue colour scheme (see Table 1). In the case study, the relation between nodes is shown by colour and size.

In this colour scheme, red corresponds with the highest item density in cluster 1 and yellow corresponds with the lowest item density in cluster 5. Furthermore the node size denotes the number of received citations (White H, 1981, V 32, P163, JAm) being the largest node in the map, while (Chen C, 2001, V34, P65, Compute) is the smallest node. This can give a great insight into the relations inside a group and between different groups.

Stage 6: Data Evaluation

The main objective of visualising the co-citation data using graphs is to highlight unusual relationships in the graphs by representing them as edges between regions of the graph that rarely occur together. In citation network, two co-authors who are drawn from groups that usually do not work together may sometimes publish together (cross- disciplinary papers). Such anomalies provide unique insights about the relationships in the underlying network. Hawkins (1980) defines an anomaly detection based graph as finding "graph objects (nodes/edges) that are rare and that differ significantly from the majority of in the reference graph nodes." Graph investigation technique permits the user to filter out nodes based on visual and semantic attributes. The method allows filtering-out nodes by their groups (colours). In addition, the method adopted in this research allows easy modification of filtering options, which may be dependent on other attributes. Each paper in the collection is associated with the authors who wrote it and the references it cites. Cluster 5 consists of papers, which covers *visualisation of literature technique*. All of the element were based on three types of literature, bibliometrics, scientometrics, and informetrics. The mutual information for cluster 5 is 0, which confirms that the elements of that cluster are not linked to other clusters and are considered as ***collective anomalies*** with respect to the entire dataset. Cluster 1 whose mutual information is 93 confirms that the elements of this cluster share common characteristics/domain area, which are *Library and information science techniques*.

In Table where cluster 1 shows high mutual information indicating higher co-citation strength, cluster 5 has a 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

We applied mutual information to detect anomalies in the context of co-citation, using the equation below:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right)$$

We computed the mutual information MI (X, Y) between two attribute sets X and Y, and only where the mutual information is greater than a threshold. We define X and Y to be dependent on: $I(X, Y) \geq \beta\mu$

Where, $\beta\mu$ is a threshold parameter set to 0.1 in our case study. Thus, for a given node we consider all pairs of dependent and mutually exclusive subsets having up to n nodes, and calculate the corresponding -values. A ratio of the form:

$$\gamma = \frac{p(X_t, Y_t)}{p(X_t)\,p(Y_t)}$$

It has been proposed as a measure of suspicious coincidence by Barlow, (1989). It conditions those two nodes X and Y should be combined into composite nodes XY if the probability of their joint appearance P (X, Y) is much higher than the probability expected in case of statistical independence P (x) P (Y). Here high values of γ are interesting as it signifies a suspicious coincidence of the events co-occurring. From Table 1 above we can conclude that cluster 1 has the highest mutual information calculation value 0.93, in comparison to cluster 5 that has the lowest mutual information calculation value 0.0. This indicates that in cluster 1 there has been a strong relationship among the nodes; however, in cluster 5 the relationship among the nodes is weak. We are interested in exactly the opposite situation, where low γ values signifies that the events do not co-occur naturally. If they are observed together, it is then treated as an anomaly. An unusually low value of the ratio suggests a strong negative dependence between the occurrences of nodes in the data. This also ensures we have seen enough cases of nodes to support the theory of negative dependence. (IL-Agure, 2016).

2. DISCUSSION

Using the 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. 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 visualised the data and examined the resulting visualization 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. Analysing 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.

In the case study, the clustering approach was used to cluster the data into groups sharing common characteristics, graph based visualization and mutual information were used to validate the approach. 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 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 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 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 theapproach developed when scaled to large data volumes and combined with semantic pre-processing, allowed us to work with noisy and inconsistent data. Mutual information supported a semantic interpretation of the clusters, as shown by the discussion of cluster 5. 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). The aim of this approach is to check data quality and any associated problems in order to discover first insights into the case studies, and detect interesting subsets to form hypotheses regarding hidden information. This approach can help to identify any

anomalies in the data, to characterise them and to understand their properties. Mutual information is a quantitative measurement of how much one random variable (B) tells about another random variable (A). In this case, information is thought of as a reduction in the uncertainty of a variable; high mutual information indicates a large reduction in Uncertainty whereas low mutual information indicates a small reduction and zero mutual information between variables.

REFERENCES

[1]   Getoor L., and Diehl C.(2005). Link mining: A survey SIGKDD Explorations, December. Vol.7 (2).

[2]   IL-agure, Z. I. (2016). Anomalies in link mining based on mutual information). Staffordshire University. UK.

[3]   Chandola V., Banerjee A., and Kumar V.(2009) Anomaly Detection. A Survey, ACM Computing Survey. 41(3). p.15.