Exploratory Mining over Organisational Communications Data

Alan Allwright[1] **John F. Roddick**[2]

[1] Defence Science and Technology Organisation
PO Box 1500, Edinburgh, South Australia 5111,
Email: alan.allwright@dsto.defence.gov.au

[2] School of Computer Science, Engineering and Mathematics
Flinders University,
PO Box 2100, Adelaide, South Australia 5001,
Email: roddick@csem.flinders.edu.au

Abstract

Exploratory data mining is fundamental to fostering an appreciation of complex datasets. For large and continuously growing datasets, such as obtained by regular sampling of an organisation's communications, the exploratory phase may never finish. This paper describes a methodology for exploratory data mining within an organisational communications dataset. A model of support for knowledge discovery is described in conjunction with a communications based concept hierarchy. This is then used as the basis for a set of visualisations. The intention of supporting visualisations in this way is to establish a sound set of requirements for the representation of communications data. The visualisations provide several interconnected representations of the data, as well as support query and drill-down into a dataset. It is suggested that this interaction with the dataset facilitates an appreciation of the data which precedes and shapes knowledge discovery. A communications analysis example is developed using the visualisations within the context of exploratory data mining.

Keywords: Exploratory Data Mining, Visualisation, Communications Analysis.

1 Introduction

Organisational communications studies[1] are often conducted in order to improve existing, or introduce new, services to users. In many organisations such services include email, WWW access, and access to distributed repositories, in addition to telephone and fax.

In order to provision the facilities for services an assessment of demand is necessary. In communications analysis demand is typically time and/or topology based. In time based studies the information traffic flow characteristics such as arrival rate, message length and service time are assessed (q.v. Jain 1991), while topology based studies tend to investigate characteristics of network connectivity such as link density and distribution (q.v. Albert & Barabási 2002).

An analysis of demand does not necessarily provide an adequate appreciation of where, in an organisational sense, people (or other business components) are creating and using applications. An assessment for the provision of services must look beyond technical measures. Neither time nor topology based analysis necessarily provide insight into the relationships between the various data dimensions. The multidimensional nature of communications data, the interconnectedness of communications components, and the sheer volume of data collected can complicate the analysis of communications systems[2].

Data mining supports communications analysis by providing insight gleaned from the total communications dataset and the provision of techniques to detect patterns and isolate data of interest. Overviews of data mining studies in the area of communications networks (Garofalakis & Rastogi 2001, Hulten et al. 2001, Julisch & Dacier 2002) offer some insight into current efforts to address these problems[3].

The approach discussed in this paper is to support an investigation focussed on a large and diverse collection of communications data. Our objective is to provide analysis in a general sense, with the proposed data mining capability placed between low level datasets and more specialised higher level mining and analysis tools. A set of interrelated visualisations are developed for this approach. The visualisations are described in the context of communications analysis, and the potential application of data mining and knowledge discovery (DMKD) algorithms. An important aspect of the visualisations is to blend communications analysis with data mining to offer new perspectives through explorations. We also describe a graphical user interface (GUI) designed to assist in exploratory data mining within a communications dataset. The objective of the GUI is to assist in representing the multidimensional aspects of communications data, and to ultimately support knowledge discovery.

The structure of this paper is as follows. A discussion on exploratory data mining is provided first in

[1] Organisational communications studies may cover the full spectrum of human communications, however, in this paper the term communications is used to represent electronic communications between organisational components. Components are used generically and include business units, people, or other uniquely identifiable parts of an organisation.

[2] Overviews of the types of networks, the range of problems, a variety of visualisations, and the general complexity of current communications networks research and analysis are provided by Keller & Keller (1993) and Dodge & Kitchin (2001*a*,*b*).

[3] The Association for Computing Machinery (ACM) has held workshops in mining network (MineNet) data in 2005 and 2006. Prior to this the related ACM workshops were in the more general area of data mining and knowledge discovery (DMKD). The principal topics of interest in the MineNet workshops are Collection, storage and access infrastructure, Network data analytics techniques and tools, and Applications to network operations and management. Within the DMKD community these relate closely with the *data understanding and data preparation*, *modelling*, and *evaluation and deployment* classifications of the CRISP-DM knowledge discovery process (Chapman et al. 1999).

Section 2. In order to establish requirements for the proposed GUI, we establish, in Section 3, what we mean by communications in the context of an organisation. In Section 4, a set of visualisations and their usage is described. Conclusions and further work are discussed in Section 5.

2 Exploratory Data Mining

Similar to mining in the general sense, the objective of data mining is to extract valuable or interesting information from a data resource. A discussion of interestingness[4] is outside the scope of this paper. However, the DMKD supporting processes of directing, filtering and isolating data are, as discussed in this paper, necessary to provision of contemporary communications analysis.

Data mining has directed and undirected (or exploratory) aspects. Directed data mining has specific goals to efficiently and deliberately extract information from the dataset that has a greater than average prospect of being of interest to an analyst. An exploratory perspective implies a more loosely framed goal. Consequently, there is greater scope for ad-hoc interaction with the dataset. What is learned through the exploratory activities can be used to inform and guide future mining activities. In either case, directed or exploratory, the resource (i.e. the dataset), and the tools (e.g. *association mining*, *clustering* and *classification*) can be the same.

Ceglar, Roddick, Mooney & Calder (2003) propose a human-centric, tightly-coupled knowledge discovery process. The process is proposed in support of the assertion that only a human can direct their enquiries to derive meaning from their interaction with the dataset. Such a directed process could only occur if conducted within a sound contextual appreciation of the dataset. This appreciation would necessarily include an understanding of the discipline and methods through which the data was obtained, knowledge of the design and organisation of the data repository, and an understanding of the range and patterns within the data.

This paper suggests a bootstrap approach where familiarity with the dataset, facilitated through human-centric exploratory data mining, fosters the development of a contextual appreciation that in-turn supports a directed knowledge discovery process. The approach attempts to presume a minimum *a priori* knowledge of the dataset. Consequently, it is suggested that the appreciation precedes, facilitates, and shapes the knowledge discovery process. However, the presumption of minimal prior knowledge of the dataset is leveraged against an expectation of a good understanding of the discipline (in this case organisational communications). A good knowledge of the discipline, in-turn, assists the analyst in framing appropriate queries within the context of the dataset. This duality forms the basis for differentiating implicit from explicit information, as discussed later in this section.

Importantly, for human-centric explorations, and in particular during their early stages, the ability to form ad-hoc queries, to filter and isolate data are critical. The issue of flexibility in the formation of queries is recognised in general in DMKD. To quote Han (Ankerst 2002):

> *Data selection and viewing of mining results should be fully interactive, the mining process should be more interactive than the current state of the art and embedded applications should be fairly automated.*

This quote also highlights a combination of capabilities very common in DMKD, namely data selection, visualisation, and interaction. These capabilities, in particular visualisation, are central to the approach proposed in this paper.

Many visualisations in the area of DMKD, and indeed within the general area of visualisation, are based upon predominantly intrinsic information. This information is essentially self evident within the context of a dataset. However, extrinsic (that is, outside of the dataset) information can be important. As an example, in association mining often visualisations are constructed to represent rules and associated quality metrics such as confidence and support (qv. Hao et al. 2001, Hofman et al. 2000, Ong et al. 2002, Rainsford & Roddick 2000). The rules and associated quality information can be derived entirely from the dataset. This tendency to consider only intrinsic information potentially overlooks valuable support from the discipline to which the rules apply. With regard to the typical market basket example, such intrinsic information includes the individual shopping items, quantity and cost. Sources of extrinsic information include taxonomies/ontologies and process models. With regard to the market basket example, extrinsic information may be derived from classes of items (e.g. cleaning products, cereals, etc.), rules for how the items satisfy the user requirements (e.g. for cleaning, for eating, etc.) and models of how the shopping is conducted (e.g. through mail-order, Internet, or visited physically). Extrinsic information considered in this paper is developed on the basis of technical characteristics of communications. Additionally, the mining process may be informed by the functions of the organisation and its communications systems.

3 Organisational Communications

All organisations are dependant upon communications for their day-to-day operation. People are increasingly becoming skilled in an ever growing range of communications technology. The ways in which people and collectively their organisations communicate are as individual as the people and the roles they conduct (El-Shinnawy & Markus 1997, Michailidis & Rada 1997). In this section a model for organisational communications data mining is developed. The model is based upon extrinsic communications factors in order to overcome limitations noted in the previous section. The model is independent of any particular dataset. Specific functions of the model are to clarify the data types and support the design of the visualisations. The model helps to scope explorations and to define what is available in the dataset. In addition, the model provides assistance in the interpretation of exploratory outcomes.

The analysis of communications within and between organisations must take into account many modes of communication. The ability to analyse predominant, or conversely unusual, modes of communication requires insight into the total range of communications for a given situation. Additionally, insight into a range of communications supports a balanced investigation across all represented communication and avoids being prematurely drawn to a conclusion. There are many techniques supporting the analysis of specific communication domains (Card et al. 1999, Keller & Keller 1993), including computer, radio or social communications. Our objective is to provide a consistent framework to collect, associate and interact with communications data from a broad

[4] A valuable survey of this topic is provided by Geng & Hamilton (2006).

range of domains. We are unaware of similar systematic approaches to support organisational communications combining the individual domain analyses. In order to achieve this, we first establish what we mean by information and develop a communications based concept hierarchy and associated characteristics.

3.1 Communications Analysis Requirements

Communications analysis can be complicated by the wide range of issues of interest to users of communication systems, the diversity of technologies and the variety of analysis tools available.

Often there are multiple users, each having requirements involving a composite of information amassed in a combination that is unique to a particular situation, such as that expressed in Figure 1.

As an example, a user, suggested by the icon centre left in Figure 1, will use technology in what they consider appropriate to a given task. A user with a requirement to collect information may correspond with other users through email, or search news and WWW repositories. The selection may be based upon an application perspective (e.g. email, news or the WWW), or perhaps a search perspective (e.g. bookmarks or the WWW). In either case, the user is presented with a relatively simple choice. However, often users do not fully appreciate that the communications resources are supported by diverse and complex communications systems. The decision as to which technology to use is moderated by a wide variety of factors including cultural and personal preferences, experience and training, corporate rules and processes, as well as the availability and applicability of technology. This situation is likely to be reflected by users through a broad mixture of requirements including what they need, what they want, what they think will do the job, and what they know about technology and services. Technically, these requirements are many, varied, and not well known or understood. Requirements from the users may be ambiguous and conflicting, and potentially users are either unaware, or unable to communicate their complete set of requirements (Gause & Weinberg 1989, Thayer & Dorfman 1995).

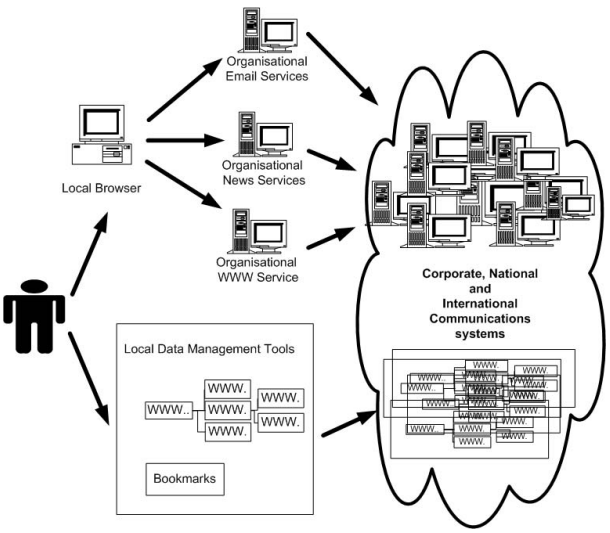

Figure 1: Computer Communications Application Diversity

In order to work within this uncertainty, we suggest a three step approach. First, we resolve what we mean by information, and then provide a set of communications systems characteristics suitable for the development of a concept hierarchy. Finally we suggest how these characteristics can support the communications analyst.

3.2 Information

A basic problem in establishing a concept hierarchy for communications is finding an adequate description of information. On the one hand, the description must cover the broad range of communication related issues encountered within organisations. On the other hand it must be technically precise enough to specify a hierarchy.

Communication studies is a broad subject covering a multitude of modes of human communication from a largely social/psychological perspective. This field is particularly useful for our work by providing a context within which a range of communications can be described within a common framework.

Within communications studies, the foundation concepts in information theory (Fiske 1990) provide a useful starting point. Unfortunately, these concepts are fundamentally contentious because they focus on the characteristics of information systems and not the consequence of the exchange. The following quote highlights this issue:

> *Shannon and Weaver's engineering and mathematical background shows in their emphasis. In the design of a telephone system, the critical factor is the amount of signals it can carry. What people actually say is irrelevant.*
>
> (Fiske 1990, p.10)

In essence, their theory is considered as a technical example within the group of fundamental communications theories. Within this paper, we accept the technical emphasis. In fact, it is the technical aspects that form the basis of our communications model. We also suggest that analysis based on technical measurements may help to frame investigations into other, less tangible communications occurring within an organisation.

In summary, the communications system proposed as the basis for Shannon's mathematical description of communications (Shannon 2001) is useful in the broader analysis of communications and is also technically specific enough to develop a foundation set of communications related characteristics. The system supports a type of information that is tangible, it flows through channels (which are carried by bearers), it has a source and a destination. A relevant communications system is shown in Figure 2. The source and destination are identifiable components within an organisation (e.g. people, organisational groups and various forms of technology). The bearer represents a medium capable of carrying information. Within the figure, the perspective, and thereby the type of query (shown as **?**), is dictated by the type of object (user, bearer, technology). This includes the type of information associated with an object, how the query should be constructed, and the types of responses.

3.3 Concept Hierarchy

Given the above communications system model, basic types of questions are as follows:

- what are the information bearers?
- how are source and destination represented? and
- what constitutes the tangibility of the information?

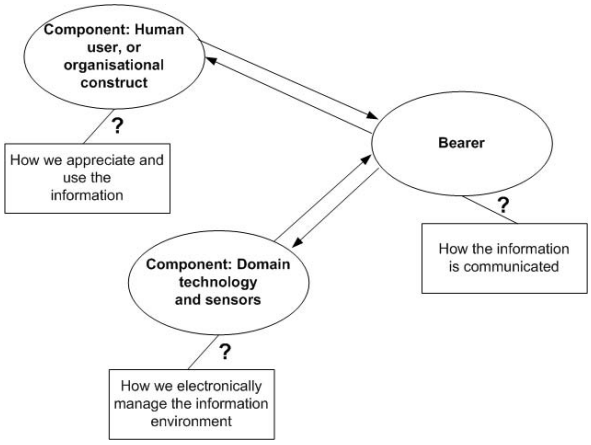

Figure 2: Communications System

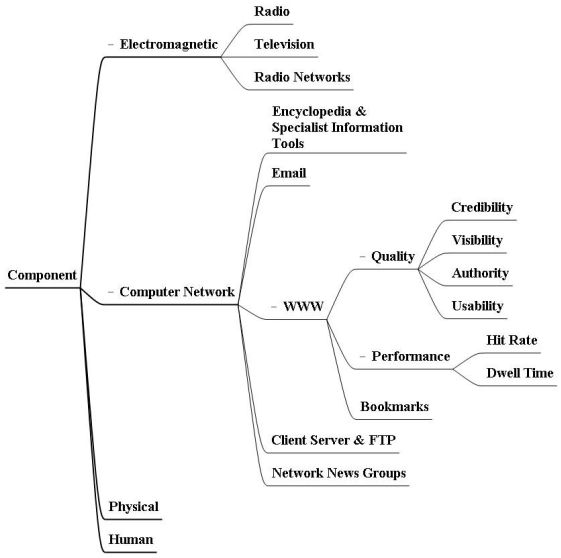

Figure 3: Example Concept Hierarchy

This information can be drawn from the communications models of Shannon (1948) or Lasswell (1948), from technical descriptions of the various technologies (such as computer networking, television or radio), and from the users requirements. Below we provide a description of the information related characteristics: bearers, addressing, and tangibility. We use the measures volume and time to represent the concept of tangibility.

3.3.1 Bearers

A detailed study of users and their target environment would normally be required to fully consider all the communications options. In this paper, a subset of characteristics regarding the bearers associated with a given user are considered, these are shown in Figure 3. The principal information bearer types are the computer network, physical, human and electromagnetic. Instances of bearers within these types (such as television, email or the WWW) would be expected to have representative systems within many organisations.

Measures, such as the availability or relevance of the WWW or television, are likely be dependent on particular aspects of an organisation. Detailed metrics, including the quality or performance of a particular medium, are very likely to be situation dependant. Consequently, representation for these measures is based upon intrinsic information.

The example concept hierarchy provides a model for the relationship between the bearer types and a subset of classes of characteristics (namely addressing, volume and time). At the level of the bearers, all share these classes of characteristics. At lower levels (e.g. specific channels), this is no longer necessarily the case. The important issue is that a communications analyst is provided with a representative class of bearers and associated characteristics for the particular aspects of a study. In the guided knowledge model of Ceglar & Roddick (2007), this would also have an interactive nature where the class of characteristics (i.e. guidance) would be generated in response to the current model of the underlying dataset (i.e. streaming).

The visualisations in the next section build upon the extrinsic information by including example intrinsic data.

3.3.2 Addressing

Within our communications model, information is treated as if it is contained within a discrete parcel while being transferred across the bearer. The characteristics of interest are either explicit (such as the parcel addressing) or evident from the bearer (such as a *television* program). In either case, the parcel has a defined source and destination. Various types of source and destination may be considered, and the communication may be point-to-point, multi-cast or broadcast. The model does not prescribe an address format, only that components are uniquely addressable.

3.3.3 Volume

The information parcel exists within a discrete space. Volume is used in the sense of the amount of information that may be contained within the parcel, and also to highlight the multidimensional nature of the communications data. There are various measures associated with the volume of information, ranging from information theoretic to simply registering the length of transmission. A parcel could have multiple measures, for example, a physical package, a number of bits, a length of transmission, or an information density. A data recording tape has all of these.

3.3.4 Time

The parcel exists in time and may have various time-stamps (e.g. when it was created, sent or detected). In addition to a time-stamp, the parcel has a duration indicating the time over which the parcel occupied a bearer.

3.4 Organisational Communications Data

Lower level computer communications information is often stored in logs at rates that present a substantial burden on automated computer based analysis. For example, WWW site contact logs (kept for caching or audit purposes) tend to be stored for days, or even weeks on machines at service providers. Each contact may include: the source and destination identifiers, page identifiers, page metrics, time stamps etc. For a busy server, such as might be found in a commercial WWW service provider or a large enterprise, this may account for a number of megabytes of text per day[5]. Longitudinal analysis may require the data to

[5]Garofalakis & Rastogi (2001), Cáceres et al. (2000) and others, discuss the network data volume issue further.

be held for extended periods of time. As an example, to determine whether to cache or mirror a group of sites, it is beneficial to monitor the traffic to the site for long enough to capture not only the steady state patterns but also transients (due to business hours, public holidays and special events).

Table 1 shows a snippet of the type of information that may be represented in such a communications dataset - information source (Src) and destination (Dest), classification details of the Bearer, and the time stamps (Time) associated with the communication. In addition, the dataset may be linked to supporting information such as the name and address (e.g. physical, geographic, business) associated with the source or destination.

Src	Dest	Bearer	Time	
			Start	End
C1	C2	Email	0900	0910
C1	XX	WWW	0905	1020
C3	C5	Phone	0915	0917
C1	C3	Email	0920	0925
⋮	⋮	⋮	⋮	⋮

Table 1: Communication Dataset Snippet

It is highlighted at this point, that even given the broad range of communications options (Email, WWW, Phone) the characteristic types (source, destination, bearer, time) are all the same. The remainder of this paper considers whether such a composite dataset can support communications analysis.

3.5 Communications Analysis

The analysis considered in this paper is largely transitive. Exploratory data mining provides an intermediate step to more specialised analysis. Important functions include, highlighting potentially valuable data, helping to weed out non-useful information, and facilitating the identification of outliers. Numerical measures include traffic aggregates, higher level metrics supporting mean value analysis (qv. Jain 1991), as well as support and confidence.

The ability to substantiate results and search for supporting data is important in the analytical process. Similarly, it is valuable to visualise a volume of information while at the same time provide insight into the detailed events that support the observation. A single transaction of considerable importance may be buried within a mass of day-to-day communications. The ability to drill down from multiple perspectives into a dataset may help to detect such occurrences. A set of interrelated visualisations, described in detail in Section 4, are designed to facilitate the observation of relationships in the gross system whilst allowing inspection of the underlying data. This is analogous to navigating from an aggregate measure (such as support or confidence) into the raw data, facilitating detailed inspection and verification.

As noted above, the ability to interactively query and form visualisations is important to DMKD. Similarly, visualisations generated on the basis of selected samples of a dataset can provide insight into a communications system. Due to the significant rate of change, and the wide variety of data types, the ability to form samples must be flexible. Experience with the dataset (such as through exploration), and a good understanding of the communications discipline should guide the formation of queries.

4 Visualisation

The transactional representation of communications records (see Table 1), in conjunction with information volume and collection frequency issues resonates strongly with areas of support from DMKD. Equally important, as discussed above, these factors highlight the need to support an exploratory perspective.

A composite exploratory data mining and visualisation based query model is described in this section. We term this composite model the *exploration map*. The objectives of the model are to:

- provide a multi-perspective representation of communications data,
- support exploration and navigation,
- focus on particular aspects of communications systems, and
- support the query of datasets.

The individual visualisations (called *organisation*, *component* and *bearer*), are described following the exploration map.

4.1 Exploration Map for Data Mining in Organisational Communications

Figure 4 shows the exploration map. This is an overview of the inter-relations between the visualisations and a dataset. The dataset comprises a list of communications related data in transactional format (as shown in Table 1). Views are generated automatically from the dataset. Interaction with the views is intended to support a comparative assessment across the communications records, through graphically guided query and navigation.

An example scenario would involve :

1. generating the views from the dataset,
2. sighting a level or type of relationship (such as highly clustered) between components in either the component or organisation views,
3. selecting the individual components to query the dataset, and
4. generating and presenting the bearer view.

The bearer view is then used to further query the dataset. In addition, the dataset may be queried directly to filter and focus attention on a subset of the data. This reduced dataset can then be used for further visual explorations.

Visualisations are presented in Figures 5 through 7. Figure 5 provides an organisation centric view, Figure 6 presents a component centric view, and Figure 7 provides a bearer centric view. Aspects of association mining, clustering and classification are described in conjunction with these visualisations. A description of additional diagram attributes that may be associated with the visualisations is also provided, at the end of this section. Figure 8 shows an overall scenario for interaction between the views and the dataset.

4.2 Organisation and Component Centric Visualisation

Figures 5 and 6 provide graph based visualisations of connections (arcs) between components. These views focus on clustering and association for records within a dataset. Components are identified by the source (C_i) or destination (C_j) address. An arc (C_i to C_j) is drawn if the pair exist on a single record.

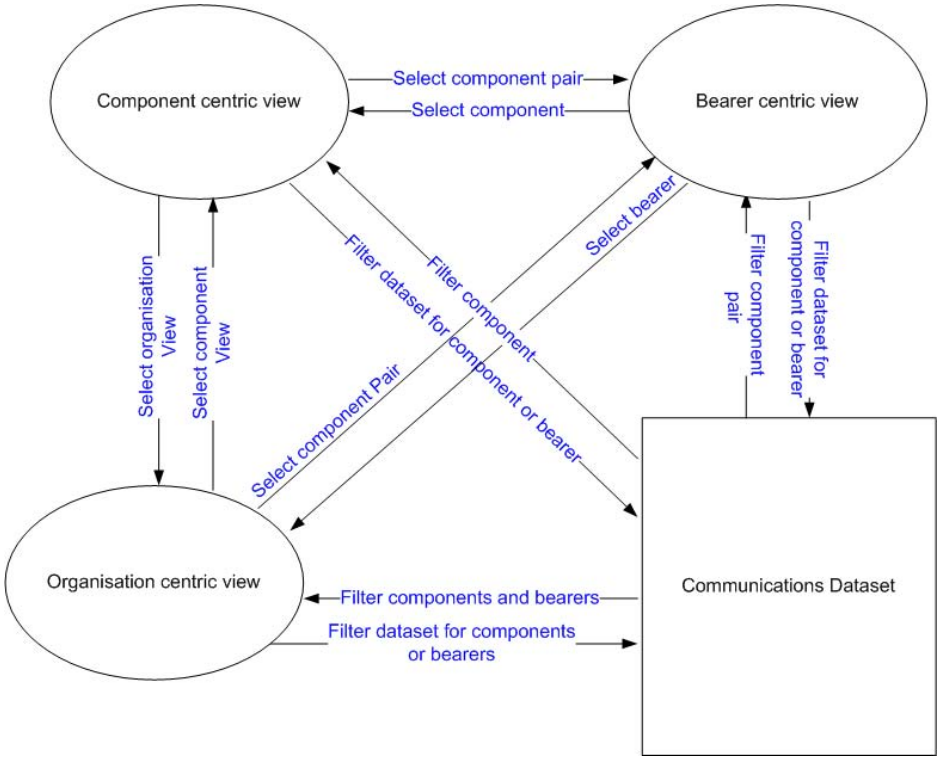

Figure 4: Exploration Map

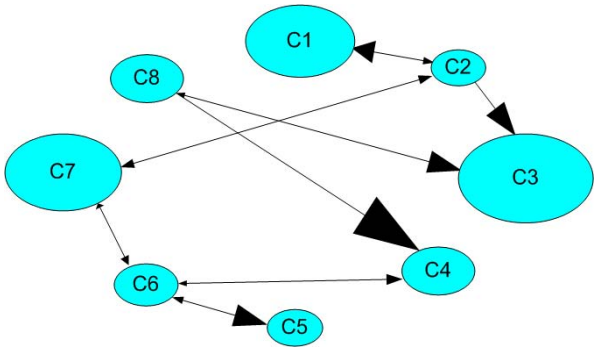

Figure 5: Organisation Centric View

Figure 5 shows the relationships between communicating components in an organisational view. The key features of this view are that it captures the complete set of components and their interrelationships as represented in the dataset. In particular, it highlights areas of concentration (many to 1) and isolation (none to 1). Data mining measures, including support and confidence, assist in substantiating any inferences.

Arrows are shown on the arcs. The direction is from source to destination, as observed in the dataset. The size of the arrows represents the relative frequency of the occurrence of the directed tuple. This provides a visual indication of the support for the two relations (i.e. $C_i \rightarrow C_j$ and $C_j \rightarrow C_i$) implicit in the connection.

The selection of an individual component (e.g. C2) queries the dataset for a specific information. The selection of an arc between two components (e.g. C1 and C2) queries the dataset and generates the bearer centric view (Figure 7).

Figure 6 provides a component centric view of a subset of the components shown in Figure 5. This view provides a means to reduce clutter and focus on specific components. It provides filtering and highlights aspects of clustering, concentration and isolation. This capability is important as the number of arcs can increase exponentially with the number of components, and may quickly obscure details of the visualisation.

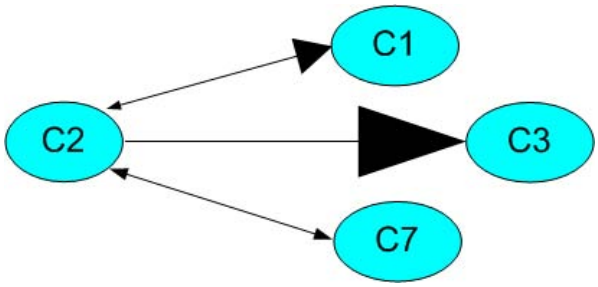

Figure 6: Component Centric View

These graph based visualisations convey a mixture of communications and data mining measures. Specific data mining tools, namely, association mining, clustering and classification, are described next. These tools are described with regard to how they support communications analysis within the context of the visualisations. Areas where additional support may be gained from the tools are also noted. In this sense, this section highlights areas of future work. An example, introducing scope for additional measures, is outlined in Section 4.4. Similarly, data mining tools are again noted with regard to the bearer centric visualisation (Section 4.3).

Association Mining : A measure of relative support is presented with relationships by varying the size of the arrows. Similarly, the thickness of the arcs could be used to represent confidence. In this sense the layout could be transformed into a Rule Graph (such as presented that by Ceglar, Roddick & Calder (2003)). Care must

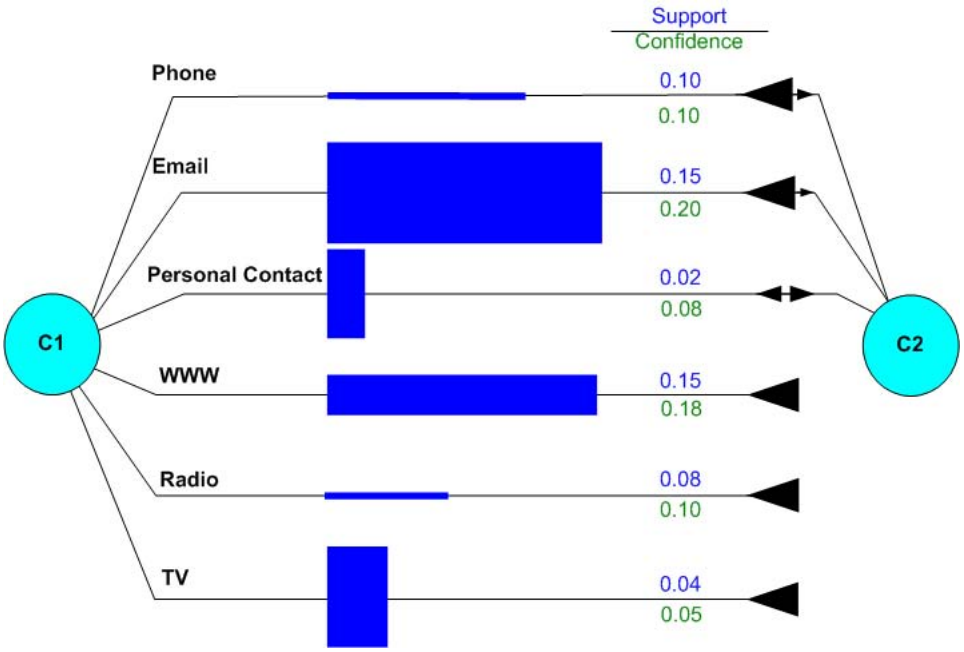

Figure 7: Bearer Centric View - showing the most significant 6 of 50 bearers

be exercised in interpreting the graph. In this instance, the length of a path (hop count) or size of an itemset cannot be assumed to be any greater than two. Inferring longer paths (or larger itemsets) could be problematic due to difficulties in adequately expressing where a given path (or itemset) starts and finishes. Sequential pattern mining applications such as INTEM (INTeracting Episode Miner and viewer) may be valuable in constructing longer path lengths (Mooney & Roddick 2004, 2006).

Clustering : Figures 5 and 6 show characteristics of connection clustering, where the number of connections attached to individual components potentially infers relationships. Other aspects of clustering can be represented within the dataset. The physical location of components could be used to construct 2D spatial clusters. The organisational or component views could also be layered over the clusters[6]. A visualisation comprising the logical clustering (based on connectivity, shown in Figures 5 and 6), overlaid upon the physical clustering (based on location) could provide important information about the efficiency of a current communications system design. For example, an overlaid perspective would support an analyst in considering whether communications are generally within or between components. Such knowledge may support network or business reengineering to more effectively utilise local services.

Classification : A high level classification scheme based on the concept hierarchy (see Figure 3) is implicit in the organisational and component views. In conjunction with clustering, a classification scheme could also be applied to highlight the organisational structure or logical business units associated with components. To extend the above example, a classification scheme in conjunction with physical and logical clustering could provide information on whether communication services are efficiently deployed to the relevant business units. An analyst, for example, could use a composite representation to assess whether business processes (such as accounting) are appropriately physically distributed.

[6]Geographic Information Systems (GIS) systems such as ArcView are designed to support this type of overlay presentation and analysis. This could be developed to provide a geographic DMKD capability as described in Miller & Han (2001).

The characteristics of organisations and components are generic, essentially only requiring that an organisation is composed of uniquely identifiable (and therefore addressable) components. The visualisations could be extended to show relationships between higher level organisations. As an example, an analyst may investigate areas of concentration and isolation between organisations in order to better appreciate preferred modes of inter-organisational communications.

4.3 Bearer Centric Visualisation

The representations in Figures 5 and 6 can easily become cluttered, and do not provide insight into the different bearers and the volume of the information. In order to overcome these limitations, the bearer centric view (Figure 7) provides an alternative perspective. The visualisation is primarily intended to provide a two dimensional histogram of the volume of information flowing between two components.

The components (C1 and C2) are the source and destination of the information. The net set of communications is partitioned across the bearers: Phone, Email, Personal Contact, etc.

The individual bars in the histogram present the relative volume of information in the dataset. The horizontal length of the bars represents the relative frequency of communication; the longer the bar, the higher the relative frequency. The thickness of the bars represents a measure of the relative quantity of information, such as the length of an email, or the time spent at a WWW site. Together, these may suggest a relative preference of a bearer by a component.

The view also shows the support and confidence results for the association $source, destination \rightarrow bearer$, (where $\rightarrow$ represents *coincidental with*), for each bearer.

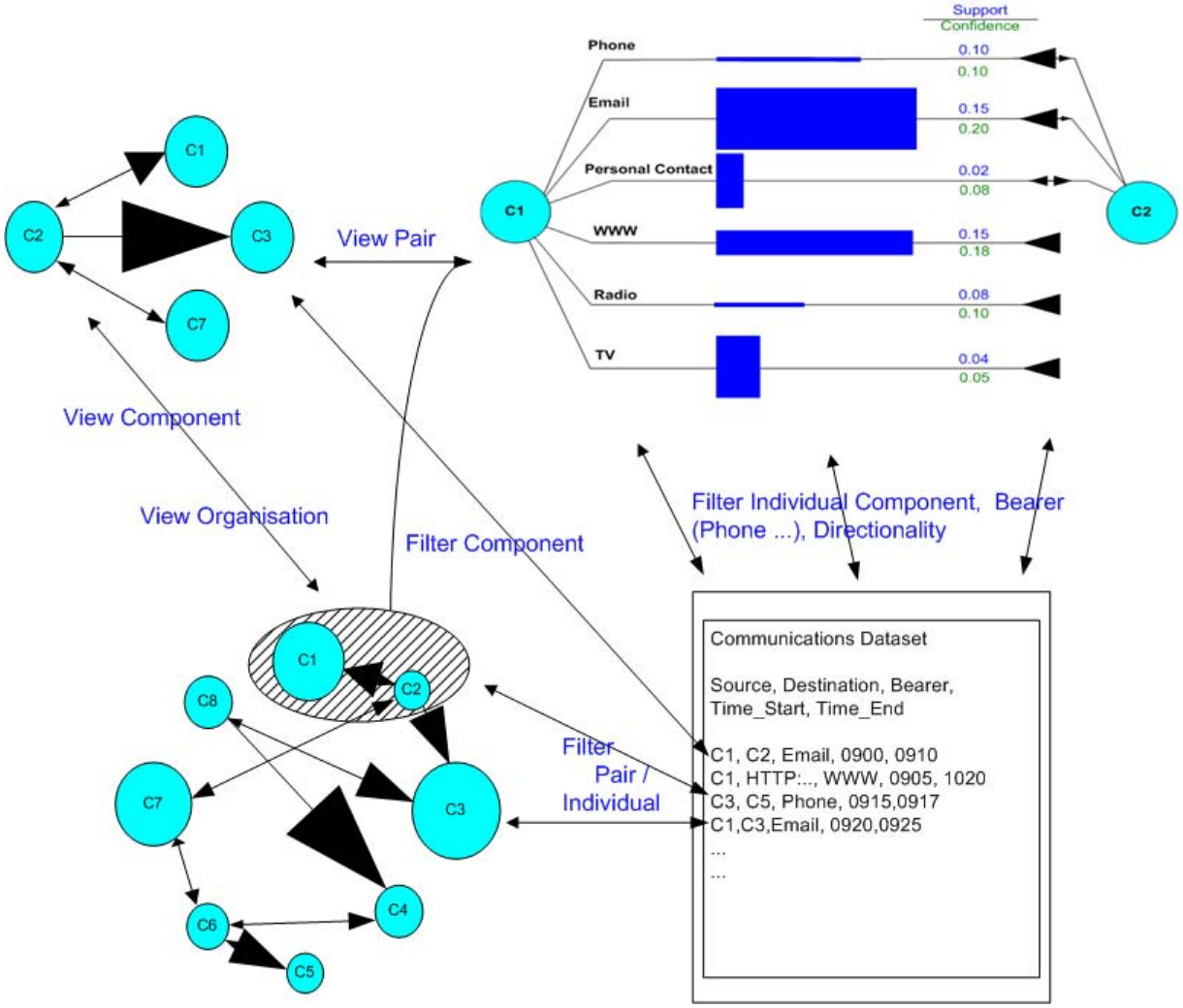

Figure 8: Integrated View

Finally, the arrows represent the overall directionality of the communications, C1 to C2, or C2 to C1. Note that broadcast media such as television, radio, and the WWW are unidirectional, and only involve the principle component (C1) as the destination. The size of the arrows represents the relative frequency of the occurrence of the directed tuple. As with the organisation and component views, the comparative size of the arrows is indicative of both the relative support in the dataset and the demand on the bearer.

The principal queries from this visualisation are the source and destination, the bearers, and the information volume. The selection of a single component (e.g. C1 or C2) or a bearer (e.g. Phone) raises either the organisation or component centric views. The component view is focussed on the selected component. If a specific bearer is selected, the dataset is filtered for the bearer before the organisation or components views are raised. Consequently, these views display only components associated with the selected bearer.

The measures, support and confidence, are introduced to blend the communications with the data mining perspectives. Further support from data mining (through the association mining, clustering and classification tools), as described next, includes additional confidence measures and higher order clustering and classification of bearers.

Association Mining : The bearer centric view is essentially an association representing the type, volume and direction of information flowing between two components.

The support measure, in combination with the volume, represents the frequency with which the combination occurs in the dataset. From a communications perspective, this provides an indication of the communications load associated with a specific bearer. From a data mining perspective, this measure provides an indication of how well the relation is supported within the dataset. An additional measure of interest would be support and confidence for the directionality. This could be valuable in locating data servers and mirror sites.

Clustering : Clustering can be applied to the bearer set, the directionality, and the importance. Within the context of components (C_i and C_j), clusters represent preferred modes of communications. In contrast to the implicit assumption of independent bearers in the association mining, a cluster of bearers may represent a preferred grouping of communications methods.

Classification : Hierarchical classification could be applied to bearers, such that sub-classes of bearers are represented. This may lessen the current restriction of only having bearers associated through components. This is equivalent to including more branches of the concept hierarchy in the bearer view. For example, rather than

Example Mapping	
Visual Attributes	**Knowledge Attributes**
Size of the nodes	Relative support for the component
Colour of nodes	Importance/priority of component
Size of Arrows	Relative frequency of the direction of distribution
Colour of arrows	Relative importance of communications

Table 2: Example Attribute Mapping

Email and WWW as two separate classes, a super class of *computer networks* could be used. Sub-classes of the computer network would include not only Email and WWW but also other forms of Internet applications, such as network news, and corporate TCP/IP applications.

4.4 Knowledge Visualisation Attributes

The size of the nodes, thickness of the lines, arrows and colours are, and could be further, used to indicate different characteristics of the relationships within the visualisations. An example mapping is provided in Table 2. Visualisation attributes are those associated with the shapes, such as lines and ellipses, presented in Figures 4 through 7. Knowledge attributes are associated with the visualisations through the concept hierarchy. Not all the attributes are relevant for all the different visualisations, either because an attribute does not contribute to the relationship, or because the data type is not available in the dataset.

The visualisations currently indicate relative directionality for communications through arrows. From a communications perspective the visualisations could be made more instructive if they included the priority or importance of components or their relationships. The relative importance of a component, as may be ascertained from an organisational chart, could be encoded into the colour or size of the ellipses. Similarly, the colour arrows could be used to represent the importance or priorities associated with relationships.

5 Conclusions and Further Work

This paper has presented a discussion on exploratory data mining within the context of organisational communications. Analysis of communications data is complicated by a variety factors ranging from ambiguity over the meanings of communications and information, to issues with collection rates and complex data interrelations. In order to work within this complexity we have developed a model incorporating problem specification, visualisation and data mining. Developing a problem specification incorporating a model for organisational communications and concept hierarchy has allowed us to form a basis for comparison across a diverse set of communications bearers. Visualisation has supported the integrated representation of communications characteristics from a variety of perspectives. This technique also appears to blend well with core data mining tools and measures. The value of an exploratory perspective in the analysis of communications data has been discussed. These elements work together to improve the potential to better appreciate organisational communications datasets.

An exploratory data mining GUI which combines visualisation with data mining tools has been described. The strengths of our proposal are the ability to use communications data from a broad range of sources, and to work efficiently through potentially very large datasets. The weakness is the inability to directly support specialised domain analysis.

The overall objective of communications analysis is to answer particular questions, or focus on a small set of issues, such as which is the preferred service, phone or email, or which WWW sites should be preferentially supported. Exploratory mining may lead an analyst to appreciate where the answer to these questions lie. Additional tools may be required for the development and presentation of results.

Several areas of further work may be considered. Enhancements to the current visualisations have been suggested, either associated with specific data mining tools or as attributes. Classification based views, such as organisational charts and product/work breakdown structures, may be necessary to provide a more complete picture. These would be included as essentially components within the exploration map toolkit.

Overall, we envisage, as further work, implementing the GUI as an interface between the mass of basic communications data and higher-level analysis tools. As an example, the GUI as described in this paper provides visually based association, clustering, and classification of representations of communications data. Tools such as ArcView may better support the final presentation. The objective is to provide mechanisms to export from a dataset into ArcView. In this way the data mining supports the context based discovery, selection and isolation of relevant data, the application of communication domain specific requirements to the mining activity, and the translation of data from communications datasets into a format for third party tools.

References

Albert, R. & Barabási, A. (2002), 'Statistical mechanics of complex networks', *Review of Modern Physics* **74**(1), 47–97.

Ankerst, M. (2002), The perfect data mining tool: Automated or interactive?, *in* 'Panel at ACM SIGKDD02', ACM, Edmonton, Canada.

Bryson, L. (1948), *The communication of ideas: Religion and civilization series*, Harper and Row, New York.

Cáceres, R., Duffield, N., Feldmann, A., Friedmann, J., Greenberg, A., Greer, R., Johnson, T., Kalmanek, C., Krishnamurthy, B., Lavelle, D., Mishra, P. P., Ramakrishnan, K., Rexford, J., True, F. & van der Merwe, J. (2000), 'Measurement and analysis of IP network usage and behavior', *IEEE Communications* **38**(5), 144–151.

Card, S. K., Mackinlay, J. D. & Schneiderman, B. (1999), *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann.

Ceglar, A. & Roddick, J. F. (2007), 'GAM - a guidance enabled association mining environment', *International Journal of Business Intelligence and Data Mining* **2**(1), 3–28.

Ceglar, A., Roddick, J. F. & Calder, P. (2003), Guiding knowledge discovery through interactive data mining, *in* P. Pendharkar, ed., 'Managing Data

Mining Technologies in Organisations: Techniques and Applications', Idea Group Pub., Hershey, PA, pp. 45–87. Ch. 4.

Ceglar, A., Roddick, J. F., Mooney, C. H. & Calder, P. (2003), From rule visualisation to guided knowledge discovery, *in* S. Simoff, G. Williams & M. Hegland, eds, '2nd Australasian Data Mining Workshop (AusDM'03)', UTS, Canberra, pp. 59–94.

Chapman, P., Kerber, R., Clinton, J., Khabaza, T., Reinartz, T. & Wirth, R. (1999), The CRISP-DM process model, Discussion paper, CRISP-DM Consortium.

Dodge, M. & Kitchin, D. R. (2001*a*), *Mapping Cyberspace*, Routledge.

Dodge, M. & Kitchin, R. (2001*b*), *Atlas of cyberspace*, Addison-Wesley, New York.

El-Shinnawy, M. & Markus, M. (1997), 'The poverty of media richness theory: explaining people's choice of electronic mail vs. voice mail', *International Journal of Human-Computer Studies* **46**(4), 443–467.

Fiske, J. (1990), *Introduction to Communication Studies*, Routledge.

Garofalakis, M. & Rastogi, R. (2001), 'Data mining meets network management: The nemesis project', *ACM SIGMOD International Workshop on Research Issues in Data Mining and Knowledge Discovery* .

Gause, D. C. & Weinberg, G. M. (1989), *Exploring Requirements: Quality Before Design*, Dorset House.

Geng, L. & Hamilton, H. J. (2006), 'Interestingness measures for data mining: A survey', *ACM Computing Surveys* **38**(3).

Hao, M. C., Dayal, U., Hsu, M., Sprenger, T. & Gross, M. H. (2001), Visualization of directed associations in e-commerce transaction data, *in* 'VisSym'01, Joint Eurographics - IEEE TCVG Symposium on Visualization', IEEE Press, Ascona, Switzerland, pp. 185–192.

Hofman, H., Siebes, A. P. & Wilhelm, A. F. (2000), Visualizing association rules with interactive mosaic plots, *in* '6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM, Boston, MA, USA, pp. 227–235.

Hulten, G., Spencer, L. & Domingos, P. (2001), Mining time-changing data streams, *in* '7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)', ACM Press, San Francisco, CA, USA, pp. 97–106.

Jain, R. (1991), *The art of computer systems performance analysis*, Wiley.

Julisch, K. & Dacier, M. (2002), Mining intrusion detection alarms for actionable knowledge, *in* '8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM Press, Edmonton, Alberta, Canada, pp. 366–375.

Keller, P. R. & Keller, M. M. (1993), *Visual cues: practical data visualization*, IEEE Computer Society Press.

Lasswell, H. D. (1948), The structure and function of communication in society, *in* '(Bryson 1948)', pp. 37–51.

Michailidis, A. & Rada, R. (1997), 'Activities and communication modes', *International Journal of Human-Computer Studies* **46**(4), 469–483.

Miller, H. & Han, J., eds (2001), *Geographic Data Mining and Knowledge Discovery*, Research Monographs in Geographic Information Systems, Taylor and Francis, London.

Mooney, C. H. & Roddick, J. F. (2004), Mining relationships between interacting episodes, *in* M. Berry, U. Dayal, C. Kamath & D. Skillicorn, eds, '4th SIAM International Conference on Data Mining (SDM'04)', SIAM, Lake Buena Vista, Florida.

Mooney, C. H. & Roddick, J. F. (2006), Marking time in sequence mining, *in* P. Christen, P. Kennedy, J. Li, S. Simoff & G. Williams, eds, 'Australasian Data Mining Conference (AusDM 2006)', Vol. 61 of *CRPIT*, ACS, Sydney, NSW, pp. 129–134.

Ong, K. H., Ong, K. L., Ng, W. K. & Lim, E. P. (2002), Crystalclear: Active visualization of association rules, *in* 'International Workshop on Active Mining (AM-2002) in Conjunction with the IEEE International Conference on Data Mining (ICDN'02)', IEEE Press, Maebashi City, Japan.

Rainsford, C. P. & Roddick, J. F. (2000), Visualisation of temporal interval association rules, *in* '2nd International Conference on Intelligent Data Engineering and Automated Learning, (IDEAL 2000)', Vol. 1983 of *LNCS*, Springer, Shatin, N.T., Hong Kong, pp. 91–96.

Shannon, C. (1948), 'A mathematical theory of communcation', *Bell System Technical Journal* **27**, 379–423, 623–656.

Shannon, C. E. (2001), 'A mathematical theory of communication', *ACM SIGMOBILE Mobile Computing and Communications Review* **5**(1), 3–55. Reprinted (with corrections) from (Shannon 1948).

Stone, G. C., Singletary, M. W. & Richmond, V. P. (1999), *Clarifying Communications Theories: A Hands-on Approach*, Iowa State University Press, Ames, Iowa.

Thayer, R. H. & Dorfman, M. (1995), *System and software requirements engineering*, IEEE Computer Society Press, Los Alamitos, CA, USA.