25

Exploiting Inter-Sample Information and Exploring Visualization in Data Mining: from Bioinformatics to Anthropology and Aesthetics Disciplines

Kuan-ming Lin[1] and Jung-Hua Liu[2]
[1]*Duke University,*
[2]*University of Leeds,*
[1]*USA*
[2]*UK*

1. Introduction

In this data-overabundant world, revealing and representing comprehensible relationships behind complicated datasets have become important challenges in data mining. This chapter presents recent achievements in applying data mining techniques to two application areas—microarray analysis and anthropology study on Wi-Fi networks—and applies visualization techniques to help integrate heterogeneous databases to obtain useful data interpretation.

As the amount of available microarray data has increased exponentially, integration of heterogeneous databases has become necessary. However, direct integration of microarrays is ineffective after normalization because of the diverse types of specific variations in experiments. This chapter reviews two approaches to overcome this issue, and introduces a cube model that combines and outperforms the two approaches by extracting information from yeast genes.

This chapter will continue on a recent anthropology study which applies data mining to visualize urban Wi-Fi networks. In the past, artists could not store and handle with huge data without database, and therefore their works typically failed to communicate with other disciplines. Nonetheless, anthropologists have explored implicit relations and viewed them in multiple ways, and they have created a number of different category principles including binary opposition, functional structure and interpretation. These multiple principles can serve as the basis to visualize human thinking and reasoning in cross-cultural and interdisciplinary study. The anthropology study to be described in this chapter will relate Wi-Fi statics in complicated fieldwork databases with easily-understood cultural phenomena.

Clear visualizations involve data aesthetics as well, which focuses on how to represent data in eye-conscious and categorized forms. This chapter will manifest that aesthetics can help expand existing data mining ideas to visual representations, with the example of combining the cube model introduced for bioinformatics data mining and the representation of regional Wi-Fi networks in spatial-temporal color charts.

2. Exploiting inter-gene information for biological data mining

2.1 Overview

Due to the increasing gap between the enormous content of sequenced genomes and the limited understanding of associating biological functions, computer assistance has been introduced to scientists for analyzing data generated from various biological experiments. In particular, high-throughput gene expression microarrays, which measure the gene expression levels in a number of experimental conditions, efficiently provide extensive amount of information for gene product examinations, and therefore become the most commonly available source of high-throughput biological data.

However, the volume and the diversity of microarray repositories, e.g. Stanford Microarray Database (Gollub et al., 2003) and NCBI Gene Expression Omnibus (Edgar et al., 2002), poses a challenge of integrating microarrays to yield more robust and accurate analysis than that on a single microarray. Direct integration via value normalization is often ineffective because of the diverse types of experiment specific variations such as lab protocols, microarray platforms, sample treatments, etc. A number of sophisticated merging methods have thus been proposed, and in this section we focus on a generalized cube framework (Lin & Kang, 2007; Lin et al., 2009), which can be combined with a variety of metrics and learning algorithms originally applied to single microarray analysis.

2.2 Related work

A representative method which merged multiple cDNA microarrays and calculated the correlations across all arrays was presented in (Eisen et al., 1998). This technique is nonetheless hard to integrate oligonucleotide microarrays because their expression values are typically not comparable. An improvement averaged the Pearson correlations over all datasets (Jansen et al., 2002), but according to (Tornow, 2003), averaging does not reflect any realistic correlation structure of the data.

Bayesian models (Bernard & Hartemink, 2005; Joshi et al., 2004) were also studied for integrating heterogeneous biological data. However, these models have not yet scaled up to large-scale data integration and mining, and applications were limited to analyzing a small subset of regulatory network.

Kernel fusion techniques (Borgwardt et al., 2005; Lanckriet et al., 2004) have been applied to biological data integration. Such techniques, however, assumed a certain underlying representation of the data (e.g., the radius basis function kernel matrix), and relied on kernelized clustering or classification algorithms (e.g., support vector machines), thus restricting the applicability and extensibility to various types of biological data.

2.3 The cube integration model

Suppose we are given a set of k microarray datasets that measure the quantity of gene products from a common set of genes. Thus, each gene of interest will be associated with k features. While conventional analysis simply merges these k feature sets with some normalization, we propose a more comprehensive integration model by considering the inter-gene relations, which is formally defined as follows:

Definition. A model M conforming to the cube framework consists of four elements (g, X, d, f) which generate a set of similarity matrices K and the corresponding cube vectors v:

- The first element $g = (g_1, g_2, \ldots, g_n)$ refers to a set of n genes of interest.

- The second element $X = (X_1, X_2, \ldots, X_k)$ is a collection of k microarray datasets to be integrated, where each microarray experiment X_m is a table of $n \times s_m$ entries storing the expression values of the n genes in g over the s_m samples.
- The third element $d = (d_1, d_2, \ldots, d_k)$ is a set of metric functions. Each metric
  $d_m : X_m \times X_m \rightarrow R$
  maps a pair of genes into a real value. The metrics can be simple functions like indication or difference function, or more sophisticated ones like correlation or kernel functions. With the metrics we associate each X_m with an $n \times n$ similarity matrix K_m by
  $(K_m)_{ij} = d_m(X_{mi}, X_{mjj})$.
  From the k similarity matrices, we thus define a set of k-dimensional similarity vectors v by considering all of the n^2 gene pairs (g_i, g_j):
  $v_{ij} = ((K_1)_{ij}, (K_2)_{ij}, \ldots, (K_k)_{ij})$.
- The fourth element $f : R^k \rightarrow \{keep, discard\}$ works as a filter on all vectors in v. Because microarray data often contain considerable amount of noisy or unavailable entries, the n^2 vectors generated need some filtering to reduce noise. For example, vectors with many low-valued correlation coefficients might be of little interest for gene expression analysis, so they should be eliminated to reduce both noise and computational time. Filtering itself can also select biologically meaningful vectors. As we shall see later, the TSP classifier imposes simple filtering criteria to identify cancer marker genes.

We demonstrate this data model via the cube illustration shown in Fig. 1 and the workflow outline in Fig. 2. As shown in Fig. 1(b), the n^2 similarity vectors can be naturally viewed as n^2 points in a k-dimensional space. Defining a metric in this space will then allow us to analyse the distances among the points and therefore find intrinsic relations among the genes.

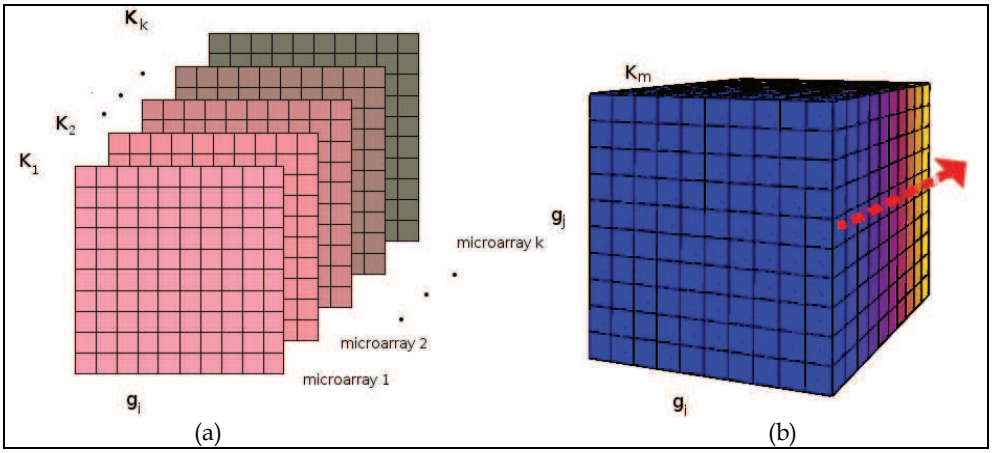

Fig. 1. (a) Each dataset X_m is associated with a gene similarity matrix K_m. Each cell in the similarity matrix is a value defined by a metric. (b) The cube constructed from the gene pairs (g_i, g_j) across the similarity matrices K_m's. The dotted line represents a k-dimensional vector.

The product from the cube framework is a subset of cube vectors labelled with gene pairs, which is demonstrated in Fig. 2. Learning algorithms can then be applied to these vectors. For example, if genes are annotated with their biological functions, these vectors can be the training data for supervised learning algorithms such as k-nearest-neighbor and support

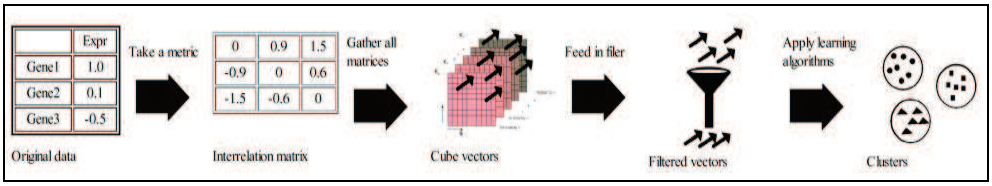

Fig. 2. Outline of the cube integration model. Each data source corresponds to one or more interrelation matrices, and all matrices form a cube. Each vector in the cube represents a gene pair. These vectors are then filtered and finally fed into any suitable learning algorithm.

vector machines. Gene clustering may also be implemented with algorithms such as hierarchical clustering and k-means. Note that kernel machines can be used to learn from the vectors regardless of whether the similarity matrices are in kernel forms, which means the cube framework is more extensible and adjustable than kernel fusion techniques.

2.4 Specialization examples and improvement

We give the outline of two integrative analysis algorithms that actually fits in the cube framework. The first algorithm, Top-Scoring Pairs (TSP) (Xu et al., 2005), translates the expression values into inter-gene comparison indicators, and then computes each gene pair's discriminative effectiveness by the distance of the center of cancer and non-cancer groups. The most discriminative gene pair is used as the marker genes for classifying new samples. Here we show how the TSP algorithm fits in the proposed integration model by identifying the four elements (g, X, d, f) of the model:

- The gene set g comprises $n = 12{,}600$ genes in the TSP paper (Xu et al., 2005).
- The input databases X is a set of profiles grouped into normal and cancer profiles. In the TSP paper three prostate cancer microarray datasets with 223 profiles were used in training, and two additional microarray datasets with 129 profiles were used in testing.
- The metric d for all datasets is the same indicator function
  $(d_m)_{ij} = 1$ iff $(X_m)_i < (X_m)_j$; otherwise $(d_m)_{ij} = 0$.
  With this metric the TSP algorithm constructs a cube with all binary entries.
- The filtering function f, which is the key of the TSP algorithm, is a ranking algorithm according to the average difference between normal profiles and cancer ones:

$$\Delta_{ij} = \left| \frac{1}{k_1} \sum_{m=1}^{k_1} (d_m)_{ij} - \frac{1}{k_2 - k_1} \sum_{m=k_1+1}^{k_2} (d_m)_{ij} \right|.$$

The gene pair achieving the highest rank is then selected as the marker gene pair.

To classify a new profile, TSP compares the expression values of the two genes in the marker gene pair (g_i, g_j). Suppose in the training data the normal profiles have higher average expression value of g_i than that of g_j. During classification the new profile will be classified as normal if and only if in this profile g_i is also more expressed than g_j.

We then show how another algorithm, second-order correlation analysis (Zhou et al., 2005), corresponds to an example in the cube framework. This study found that doublets of the same function may have moderate overall (first-order) correlations but high second-order

correlations which were not considered by conventional correlation analyses. Again, we identify the four constructs of the cube model as follows:

- g is a set of 2429 genes from budding yeast *Saccharomyces cerevisiae*. All genes are annotated based on Gene Ontology (Ashburner et al., 2000).
- *X* consists of 35 cDNA microarrays and four Affymetrix ones. Unlike the TSP algorithm, in each microarray a gene corresponds to a group of profiles.
- The metric *d* in each microarray is jack-knife correlation, which takes the leave-one-out Pearson's correlation coefficient with the minimum absolute value. The use of jack-knife correlation effectively reduces the number of doublets in the filtering phase.
- The second-order correlation analysis study poses several constraints on selecting doublets to overcome computational difficulties. First, only the gene pairs from the same functional category are included. Furthermore, they consider only the genes where at least eight expressions are available in all microarrays. Finally, a gene pair is defined as a doublet if at least eight and at least a quarter of the correlation values are greater than a cut-off value $\tau=0.6$. The selectivity of the filter *f* is low, as only 5142 doublets pass the filter.

The doublets thus selected are then clustered using TightCluster (Tseng and Wong, 2005) with the similarity metric in the *k*-dimensional vector space being correlation again, which gives the name of the second-order correlation analysis. In general, however, the distribution of the points in the vector space might be captured by other metrics like Euclidean distance or normalized inner product. According to the experiments in (Lin & Kang, 2007), the second-order correlation analysis could be improved by applying different metrics or learning algorithms to the filtered vectors. As shown in Table 1, the accuracy of biological function classification was improved by simply changing the metrics or the clustering algorithm.

Metric	Clustering algorithm	Inclusion accurary
correlation coefficient	TightCluster	72% (Zhou et al.)
correlation coefficient	hierarchical clustering with complete linkage	90% (Lin & Kang)
1-norm	hierarchical clustering with complete linkage	93% (Lin & Kang)
1-norm	c-means	89% (Lin & Kang)

Table 1. Functional homogeneous group inclusion rates over the 100 tightest clusters.

2.5 Discussion

Although the two algorithms in the previous subsection seem to differ in both the analysis models and the target applications (i.e., gene functional classification and disease prediction), they both utilize inter-gene information and can be well described under the cube integration model. Therefore, the cube model can server as a general integration framework that provides an easy and efficient way to implement an effective microarray integrative analysis framework.

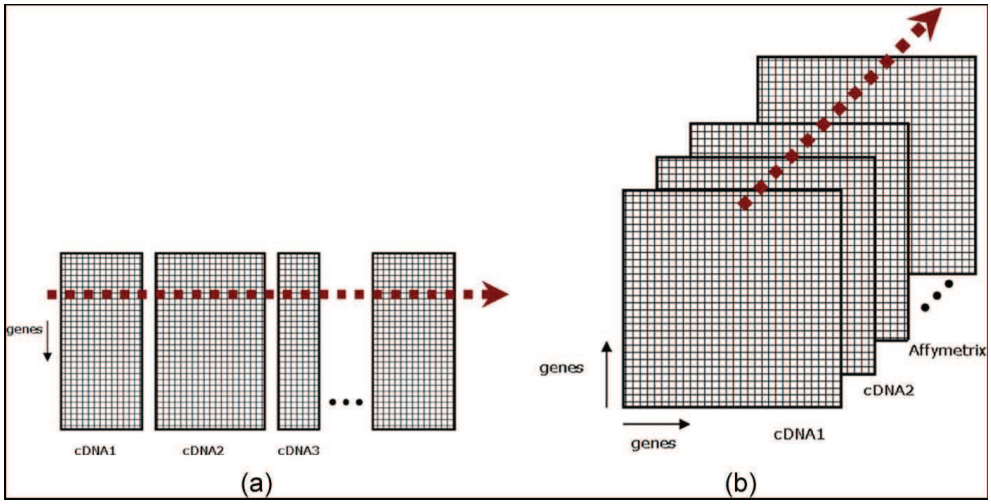

Fig. 3. (a) Illustration of conventional microarray merging technique. The microarrays are concatenated to form a large table. The dotted arrow represents the attributes of a gene. (b) Illustration of the cube framework. The similarity matrices generated from microarrays are piled to form a cube. The dotted arrow represents the attributes of a gene pair.

To summarize the intrinsic differences, a visual comparison between the conventional merging technique and the cube framework is shown in Fig. 4. The cube model is more flexible than the concatenation model and has several immediate advantages. First, the construction of the cube vectors is independent of the analysis, so we can apply any clustering or classification algorithm as long as the pair-wise distances between gene pairs are defined. Also, since now the complexity of each microarray is hidden in a matrix, in the analysis phase there are no normalization or feature selection issues, as long as the number of microarrays is not too large. Moreover, domain knowledge for a specific microarray experiment may be applied by designing the metric function for the dataset to improve the performance of subsequent data analyses. In sum, the integration model not only reduces the effects of experiment specific factors, but also captures vital information of inter-gene relation in biological processes.

3. Exploring urban Wi-Fi landscapes in anthropological fieldwork data mining

3.1 Overview

This section introduces a Wi-Fi (wireless technology for connecting to the Internet) landscape research project which explains how anthropologists study and analyze cyborg (cybernetic organism) identity in large scale Wi-Fi data collected in several cities via contextualized data mining. Conventionally, anthropologists live and study in small scale settlement to observe their participant' activities, and hence they need public and privileged statistics data as complementary data to study urban life. Such methodology however introduces a conflict between participant observations and statistic analysis on data interpretation. While participant observations focus on the data in social and cultural context, statistics concern about the data in related variables. For example, some anthropologists consider crime represents one way to solve social conflicts, but statistics

may suggest that crime relates instead to unemployment rate. The disagreement leads to develop new types of contextualized data mining methodology to obtain data and produce statistics from fieldwork.

This fieldwork for the urban Wi-Fi landscpe study explores five cities—Taipei in Taiwan, Chicago and New York in USA, Hong Kong in China, and London in UK—where Wi-Fi is widely adopted to construct the citywide networks. Low cost and easy installation are major factors for cities to promote Wi-Fi as their fundamental digital infrastructure. Beyond technology and infrastructure, Wi-Fi also relates to our habitus (Boudieu 1977) and behaviors, at least in a number of countries. Habitus shape Wi-Fi users as identifiable groups and Wi-Fi access points as the main objects in such groups. Users need Wi-Fi access points to connect to cyberspace and construct their identifiers in the social and hardware networks. Wi-Fi access points build their society, which is similar with the *house society* concept in anthropology, a society whose basic social unit is a *house* (Lévi-Strauss 1982). Here the term *house* can refer not only to physical building, but also to abstract concept.

The relation between Wi-Fi access points and houses is of conceptual metaphor (Lakoff and Johnson: 1980), where they share common properties categorized into *qualities*, *entities* and *functions* (Ahrens 2002, 2010). *Qualities* mean attributes of objects. For example, house and Wi-Fi access points share the *tangible* quality. *Entities* refer to the parts, such as houses' pipe and Wi-Fi access points' antennas. Finally, Wi-Fi access points provide connections and houses provide dwelling places, which explain the *functions* of the two objects.

Above all, the house society theory provides an innovative approach to examine Wi-Fi data in social aspects beyond the amount and distribution of Wi-Fi access points. This is significant in expanding data mining applications to include social contexts as variables in data analysis. In this section, we will introduce the development of Wi-Fi in cities, and argue how conceptual metaphor and *house societies* concept highlight cultural or social context to place Wi-Fi statistics for data mining.

3.2 Literature review

3.2.1 Fieldwork

Fieldwork is the main difference on methodology between anthroplogy and other disciplines. To obain first-handed information, anthropologists stay for a long period in the arena of their study to proceed their research. In particular, to acquire direct and clear data, anthropologists participate local events and activities to observe the society and ask informants questions, called *participant observation*. For anthropologists, *behaving and thinking like a native* is the ideal goal, and *native's point of view* is the basic requirement for explaining and interpreting cultures. Because participant observations are made by directly inquiries, anthropologists applying such apporach can only study small-scale settlement (Evans-Pritchard 1940; Malinowski 1922). Hence, if researchers want to study cities, they need to refer to statistics from government or companies (Low 1996). However, statistics is not easy to integrate into traditional anthropology methodology. To bridge the gaps, in this section conceptual metaphor will be applied.

3.2.2 Wi-Fi development

Wi-Fi, a trademark of Wi-Fi alliance, means wireless technology based on IEEE 802.11 (Wi-Fi Alliance 2010). Wi-Fi has changed the structure and meaning of Web. One decade ago when Wi-Fi were not available, Internet users relied on wired connections at home, offices

or cyber cafes to surf online because they are forced to stay at a particular position to connect to the Internet via a network cable. Web surfers were literally points in a web connected via lines. With the new Wi-Fi technology, access points (AP) emit short-distance wireless signals for Internet connection and hence create a temporary cyberspace for Internet connection. As a result, Wi-Fi users are more like residents in houses or apartments and they are connected by spaces rather than lines.

According to geographical surveys (Torrens 2008; Schmidt and Townsend 2003; Jones and Liu 2007), Wi-Fi is now very popular as millions of them are distributed in the world. These studies, however, focused only on the distribution on geographical space and applied statistics to depict Wi-Fi maps. In contrast, in this chapter we will investigate Wi-Fi distribution statistics in social space and cultural context.

3.2.3 House societies theory

Ubiquitous Wi-Fi access points become a new type of houses which locate residents/users in the physical world. For example, the search engine giant, Google, was criticized about its Wi-Fi scanning in Australia, knows so much about individuals, that is driving around and taking photos of every street in Australia, is collecting data that could enable it to physically map that information to a physical street and presumably a physical house. Besides Wi-Fi access points, Wi-Fi-equipped devices also transform human beings to cyborgs (Haraway 1991) dwelling in urban Wi-Fi spaces. To help analyze prevailing urban Wi-Fi access points with visible properties as devices and invisible information as houses, the house societies theory in anthropology is introduced in research.

Anthropologist Claude Lévi-Strauss (1982) employed *house* which contained tangible and intangible properties to study kinship systems in different societies, e.g. American natives, European nobles and medieval Japanese societies. Beyond blood relationships, material objects (e.g., wealth, space) and immaterial properties (e.g., title, fame, power) contribute to construct identities in the same house. As a result, house societies theory posits that a house is an elementary unit where economical and political activities occur, and that these interrelated activities shape house members as a special and identifiable group. This theory together with conceptual metaphors can bridge Wi-Fi access points and houses.

3.2.4 Conceptual metaphor

The connection between houses and Wi-Fi access points can be constructed by metaphor. Metaphor means applying one thing to describe another thing. Conceptual metaphor was proposed by George Lakoff and Mark Johnson (1980). They claimed that conceptual metaphor is based on culture background so we can comprehend the social/cultural relationship behind the link of diverse category of objects. They defined the item to be described as the *Target Domain (TD)* and the item to explains TD as the *Source Domain (SD)*. In this chapter, house is the SD and Wi-Fi AP is the TD.

Houses and Wi-Fi APs are distinctive matters, but conceptual metaphor bridges these two via mapping principles. Mapping principles are the reason which people connect objects of separate domains. The mapping principles concern about shared aspects on both objects. Kathleen Athens (2002) addressed three classifications of aspects—*entities*, *functions*, and *qualities*—to analyse conceptual metaphors and explore mapping principles. *Entities* mean parts such as windows of houses; dwelling is one purpose of houses and it is also the *function* of house; *qualities* are applied to describe attributes like "concrete" or "huge." The

three classifications are the variables in statistics and they relate to contextualized objects like houses and Wi-Fi access points. The values of the three variables can give weights in the mapping principles and represent the cultural/social meaning of object-centered study.

3.3 Methodology

In the following table, the three classifications of "Wi-Fi AP is House" metaphor and the examples are listed. The three classifications of conceptual metaphor provide clear criteria to describe the features of Wi-Fi access points.

	House	Wi-Fi AP	Commons
Entities	bricks, cements, windows, doors, addresses, location, ownership, residents, space, house names	box, antenna, signal, location, ownership, users, space, Wi-Fi AP names, membership, appearance	location, community, ownership, users, membership, space, name, appearance
Functions	connection, living, shelter, resting, sharing, storing up, locating person in society, providing resource	connection, transferring information, identifying members, dwelling, accessing Internet	identifying members, resource, connection, dwelling
Qualities	colors, sizes, urban, rural, owned, rented, public, private	colors, sizes, subscribing, private, public, urban, rural	colors, sizes, ownership, public, private, urban, rural

Table 2. Three classifications of the "Wi-Fi AP is House" metaphor.

The above table shows that houses and Wi-Fi APs have commons in specific items and the mapping principles as follows.

- "House and Wi-Fi AP" contains tangible and intangible entities.
- The main function of "House and Wi-Fi AP" concerns about locating people in networks/societies.
- Qualities of "House and Wi-Fi AP" can be private, public, or restricted.
- Wi-Fi AP is the materialized and spatial identity in conceptual metaphor of house.

While mapping principles provide a basis for statistics, fieldwork constitutes the collection of Wi-Fi access points, our basic data sets. This fieldwork was conducted in London, Chicago, New York, Hong Kong and Taipei. Instead of recording all Wi-Fi AP data, the fieldwork focused on how Wi-Fi networks are shaped in different cities. Although there are some organizations providing Wi-Fi distribution maps, there still exists some restrictions such as:

- Data attribute: Pure Wi-Fi machine data cannot give us culture, society, economy and environment information in the cities.
- Legal issue: Commercial companies such as Skyhook collect Wi-Fi positions by their stuff and only allow limited license to query the data from their user interface; full access to their databases is not allowed. Open/Free map groups such as wigle.net collect data by the participants, but the group owner rejects the query request because they consider the use may be unrelated to the purpose of open/free map project.

To collect Wi-Fi hardware information, a Windows batch file run in laptops and an iPod Touch commercial software, *WiFiFoFum*, are employed to automatically record Wi-Fi information while walking around the cities. Besides automatic collections by both devices in different areas of the cities, related culture and environment were also observed and notes were made.

3.4 Result

	New York	Chicago	London	Hong Kong	Taipei	Total
#Records	18428	5196	18743	84437	102047	228851
#Access points	11223	3459	3923	16740	13125	48470
Fieldwork duration	Jan 2009 – Feb 2009	Oct 2008 –Dec 2008	Oct 2007– Mar 2010	Apr 2009 – May 2009	Jul 2006 –Jul 2010	Jul 2006 –Jul 2010
Area	Manhatta n & Queen	Midtown & Northern Area	Zone 1 & 2	Hong Kong, Kowloon, Mong Kok, New Territory	East Area	N/A
Max # recurrence	42	97	178	210	1138	N/A
Average # recurrence	1.64	1.5	4.8	5.0	7.8	N/A

Table 3. Fieldwork statistics of the five target cities.

In this project, we have collected 48,470 Wi-Fi access points in five cities. Table 3 shows the statistics obtained. The total number of access points of this project is far lower than Kipp Jones and Ling Liu's work, where more than 5,600,000 access points were collected in USA by Skyhook (Jones and Liu 2006), and Torrent's work, where 500,000 access points in Salt Lake City in USA were collected (Torrent 2008). The main reason is that the collection approaches differ. Skyhook collected access points by cars with special Wi-Fi detecting facilities, and Torrens searched access points by foot, bicycles and cars with a special signal-detecting device. They aimed to include all access points in every area. Our study instead concerns about the behavior of general users perceiving Wi-Fi access points as (via conceptual metaphor) conceiving houses in their daily life. Thus, to reflect general user experiences in the statistical distribution of Wi-Fi access points, in this project walking, buses, trams, trains and metro trains serve as the main transportions. Furthermore, data were collected by personal laptops and personal mobile devices instead of specific W-Fi detection devices. In contrast to the previous two studies, the statistics context in this study relates to persons in cities rather than to urban infrastructure.

There are several additional findings in the statistics:

- Hong Kong has the highest number of Wi-Fi access points, while Chicago has the fewest ones. As we only stayed at Hong Kong for one month to collect data, we can infer that the density of Wi-Fi access points is the highest in Hong Kong.
- The maximun recurrence is 1,138 in Taipei and the minimun is 42 in Taipei. The result may be caused the familarity of the authors with the city. We can assume if users are familiar with the city more, they have more fixed routes to explore the city.

- The average recurrence presents that Chicago (1.5) and New York (1.64) have the similar average. The average of Hong Kong (5.0) and London (4.8) are also close. This phenomenon indicates that similar historical and cultural background may shape similar Wi-Fi distribution.

Two Wi-Fi access point quantities are reported for each city in Table 3: one is *records* and the other *access points*. The former is the total records without removals of the recurrence of Wi-Fi access points and the latter only counts distinct Wi-Fi APs. Unlike other researchers who filtered out identical access points in the databases to conduct the statistics of Wi-Fi APs in geographical distributions, we retain recurrences to observe two additional properties: frequency and routes. For example, commuters in Taipei shows obvious recurrence location sets, as the same access points always appear in daily routes. In this case, density of Wi-Fi APs only reflects the spatial distributions, but frequency is more meaningful. Because most users have similar places and routes to access Wi-Fi Aps in home, office, cafés, pubs and restaurants, higher frequency can mean more fixed routes.

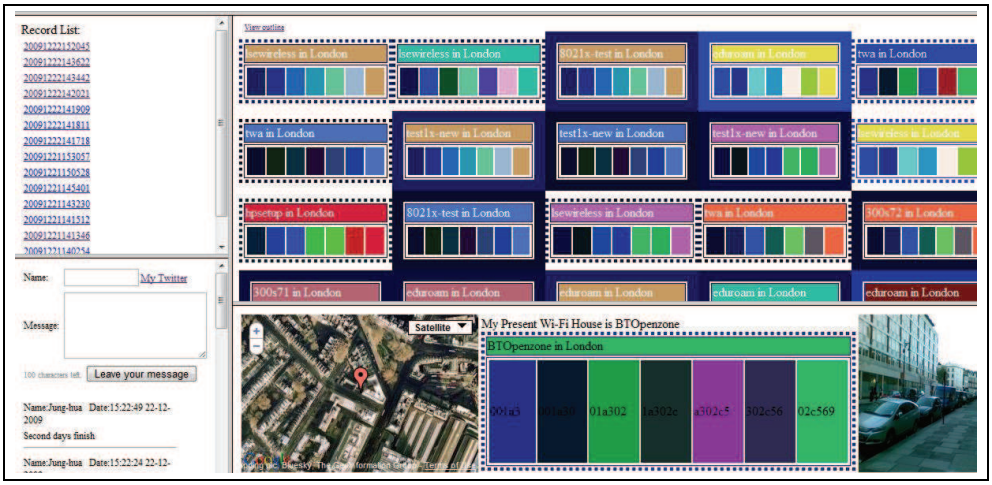

Fig. 4. Sample Wi-Fi records in London containing the location which the data are collected and uploaded, one streetview picture and the color chart of BSSID (Basic Service Set Idendtifier, Wi-Fi APs's unique identifier). The top left part is the record list and the below is the guestbook which link to this project's twitter. Converted color charts of BSSID of a series of Wi-Fi APs are presented in the top right. The bottom of right column is composed of map, one BSSID's color chart and one streetview picture. Color conversion rules are explained in the aesthetic part of the article.

We can see from Table 3 that Wi-Fi access points are popular in all of the five cities. In London, many houses have their own Wi-Fi access points and telecom company BT has installed many commercial Wi-Fi hotspots in most public places. In addition to static access points, Wi-Fi access points also appear on particular trains and coaches, as well as airplanes in London, New York and Chicago. On one hand, most research and advertisements about prevailing Wi-Fi access points emphasize Wi-Fi's easy access, wireless attributes, and commercial business. On the other hand, artists, sociologists, and geographers consider Wi-Fi as a social movement of digital citizenship in our society. They offer an all-encompassing

social view to explain and interpret how Wi-Fi technology shapes Internet life in a new, wireless way. Their concerns tend to explore public lifestyles, but Wi-Fi also plays an important role in our private lives. As mentioned, *house* is the keys to understanding the effect of Wi-Fi on the connection between private and public spheres.

3.5 Discussion

Ubiquitous Wi-Fi hotspots not only provide citizens a convenient vehicle to access the Internet but also break the boundary between public and private areas. Wi-Fi users can access the Internet in cafés or train stations in a similar manner to accessing it at home. In addition, Wi-Fi devices are widely adopted in many houses, whether they are supplied by an Internet Service Provider (ISP) or purchased by users. Wi-Fi devices are like tubes that produce heat; however, they are different in that Wi-Fi devices use wireless signals to connect two different worlds – the real world and cyberspace. Internet connections project physical users into cyberspace where they become another identity. In this sense, cyberspace is a metaphor for and a symbol of our real lives and Wi-Fi network devices are important media that create and embody these metaphors and symbols.

In contrast to telephone/modem cables, Wi-Fi devices expand Internet connections from a socket-like tap to a wireless signals. The process is similar to how energy resources create heat that is accessible to every member of the house, no matter where the member is. The main difference is that, while house members can access directly water from taps, heat via gas or electricity, and television reception through television signals, they need extra Wi-Fi equipped devices to access the Internet even though a wireless signal exists in the house.

Through connecting to the Internet via network cables, telephone lines, or modems, computers can be viewed as extensions of these devices, as house members use them in the same way as watching TV programs on television or washing hands by turning on a water tap. Under these circumstances, members use their own body parts to interact directly with the world around them. In contrast, if house members adopt Wi-Fi to access the Internet, their computers are no longer required to be fixed in particular positions. In other words, computers *escape* from modems and telephone cables and are no longer extensions of them. Instead, computers and other mobile devices can be viewed as extensions of the human body, expanding human beings' senses and abilities in order to explore the other world, *cyberspace*.

When we examine wireless maps of broader districts such as street blocks in cities, physical houses may also be defined by Wi-Fi devices as their signal coverage. This correspondence again bridges the maps of real world and cyberspace. Wi-Fi devices create network names that appear on users' computers. Wi-Fi networks are named by either manufacturers or owners, and these names are used to identify ownership of the access points. Using Wi-Fi network names, members can access the Internet via these devices. The members who are granted access know each other because Wi-Fi devices can cover only limited ranges, within a house for example. On the other hand, Wi-Fi networks are a metaphor for houses because Wi-Fi network names, like other kinds of names of houses, can be considered intangible property of houses.

Wi-Fi access points, as Claude Levi-Strauss's *house societies*, shape members into special and identifiable groups. A Wi-Fi network is both a shared property among house members and a resource for members to access the Internet. This resource is limited and restricted to particular target groups, namely family members and house members. It is worth pointing

out that house members are different from family members in that house members may be family members, flat mates, tenants and owners, etc. Because Wi-Fi networks are representations of and metaphors for a house, house society theory evokes the consideration that Wi-Fi devices in houses are viewed as more cultural and social than technological, for their influence in daily life shapes and confirms members' relationships.

4. Aesthetic view of inter-sampling for house-like Wi-Fi access points

4.1 Overview

In the previous two sections, we studied two types of data mining targets. The microarray cube framework was applied in gene analysis, and conceptual metaphor between house and Wi-Fi was applied in anthropology investigation. Both studies provide indirect methods to find innovative and interrelated information in huge and complicated data. From aesthetic point of view, such apporach is similar to how Claude Monet applied water and flowers to paint the inexistent light on his work *The water lily pond*, according to French philosophist Alain Badiou (2006). Our two studies utilize simple visible forms, cube representation and metaphoric house respectively, to present sophisticated data structures, gene microarray network and urban Wi-Fi landscape. The new forms contain and reproduce a series of rules, visual elements, orders and patterns.

Beyond algorithms, aesthetics offers data mining a new way to construct its framework to analyse the data. Aesthetics is one kind of cultural form and symbolic system, and in this sense Lin's microarray cube is biologically and Liu's metaphor is anthropologically cultural form and symbolic system. Therefore, in this section we will apply data mining to visualise and materialise these data sets and create Wi-Fi artworks in aesthetical framework. We hope that the synthetic approach will help data mining to expand to cultural/social spheres and to organize and visualize humanity data.

4.2 Literature review

Aesthetics is one of the main topics in art and art history. Different artists have their individual techniques and tastes to construct their artworks. According to Theodor Adorno (1997), a famouse philospher on aesthetic theory, the emprical world is mediated to art with the aesthetics. Aestheitcs is a series of patterns and symbolic forms to represent the world. Contemporary artists' artworks, especially generative art and net art, reflect this concept.

Generative artworks are created by one or a series of algorithms which produce a repetative patterns to present artists' interpretation for what they perceive. An example is Jared Tarbell's *Substrate*, shown in Fig. 5, which creates sets of lines to construct rectacles with algorithm to compose a visually and rythematically aesthetic world.

Net artists also create rule-based artworks with Web browsers. For example, Lisa Jevbratt's famous art project (shown in Fig. 6) converted IP addresses into five interfaces (migration, random, every, excursion, and hierarchical). This project aims to depict web into a visual language with different patterns to mediate IP addresses in physical world to imaginary world.

Above all, aesthetics is not merely a subjective and arbitrary taste for beauty; rather, it is a way to analyze, realize and represent our world in particular patterns to dig out the relations and information behind complicated data. This section will convert Liu's Wi-Fi data with Lin's cube model to create artworks to mine data with aesthetics.

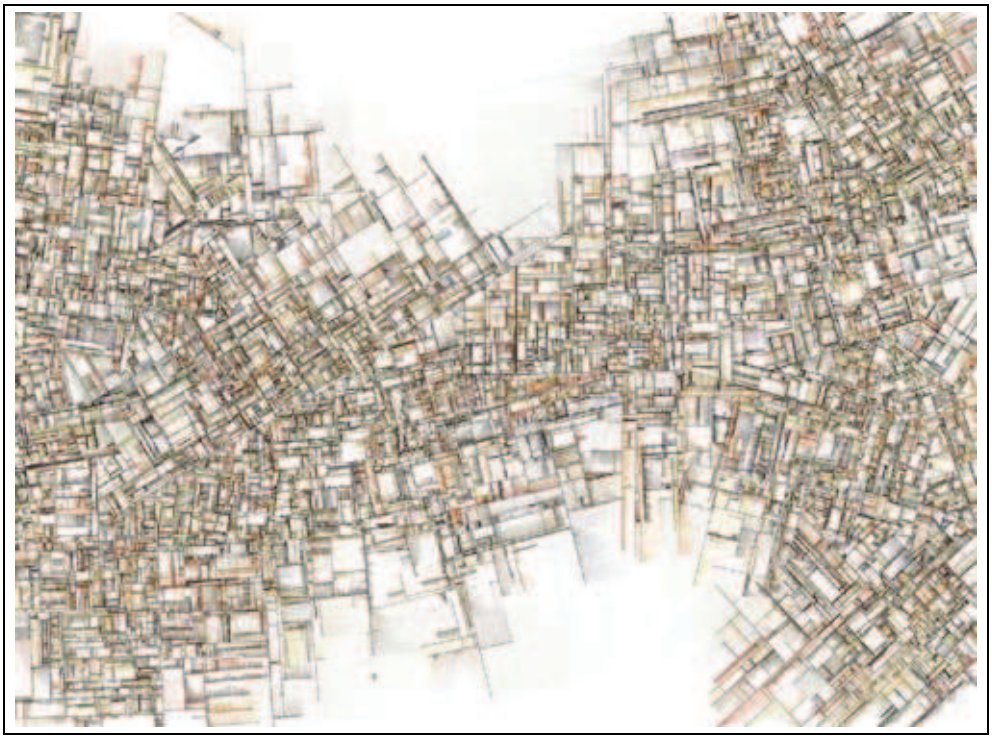

Fig. 5. Jared Tarbell's *Substrate,* a generative artwork.

Fig. 6. Lisa Jevbratt's *1:1* project.

4.3 Methodology

To convert urban Wi-Fi landscapes into artworks, an element to bridge the two domains is necessary. We thus introduce BSSID, the unique identifier of Wi-Fi access points like house addresses. A BSSID is a 12-digit hexadecimal code, composed of two parts. The first six digits is the vendor's code and the last six digits is the serial number in the factory.

Inspired by Lin's cube mode, BSSID is mediated to color charts to present the codes (see Fig. 7). For connecting to database and making artworks accessible, the converted color charts are presented in Web pages with PHP (Hypertext Preprocessor, a script language). Web pages show seven colors per BSSID according to HTML's 6-digit hexadecimal coding in BSSID through the following procedure. The first webpage color code is created from the first to the six digits of BSSID, the second color code from the second to the seventh digits, and so forth. The shifting translations reflect the coexistence and interaction between the vendor and the individuality of the Wi-Fi access point. Such transitional charts can be viewed as genes of urban Wi-Fi systems, thus similar with a gene set in Lin's microarray study. Furthermore, artworks composed of charts could provide us a new way to mine data to realize the difference among urban Wi-Fi landscapes.

Incorporated with Liu's house metaphor connection, color charts are then shaped as a *house* image (Fig 5.). The image is composed of the following three parts: (1) the bottom part, which is the color charts computed by the aforementioned method; (2) the top part, where the Wi-Fi access point name and area appear in the color of the final 6-digit color (same as the rightmost color in the chart); and (3) the border that surrounds the top and bottom parts, which is the first 6-digit color (same as the leftmost color in the chart). If the Wi-Fi access point requires authorization for access, the border will be solid; otherwise, it will be dotted.

Most physical houses in cities were built by a few building companies, and hence they usually seem similar to one another. To house members, however, their houses are unique, and our Wi-Fi access point color charts can reflect such personal uniqueness. Therefore, the color transitions in Wi-Fi networks contain both personal and public access.

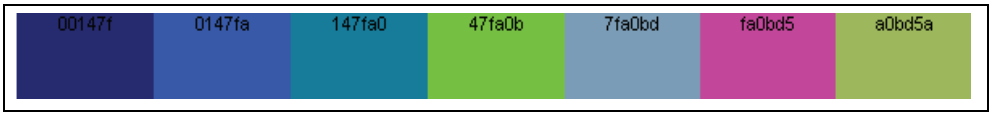

Fig. 7. The seven-color chart of the BSSID *00147fa00bd5a*. The first (leftmost) color encodes the first six digits *00147f* into RGB color code #*00147f*, which corresponds to dark blue.

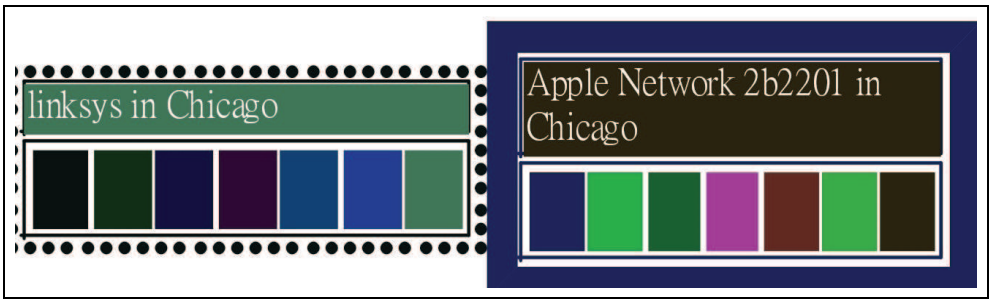

Fig. 8. House images of two Wi-Fi access points.

4.4 Result

The algorithm of this series of artworks converts BSSID to color and arrange them according to timeline. The order is from top to bottom in the same row from left to right. The arrangement is similar to stripping Lin's cube representation to long patchs, providing a new perspetive to view the data sets. The border represents not only the Wi-Fi access point's degree of openness, but also the color generated from the vendor's code. Since authorization and encryption functions are provided by the manufacturer, borders in metaphor can correspond to the walls of physical houses constructed by building companies.

The charts of London and Chicago are shown in Fig. 9. In both cities there are obvious parts which is full of Wi-Fi access points of solid border, yet Chicago seems more open than London according to the color distribution. Wi-Fi charts in New York and Hong Kong (Fig. 10) show similar patterns, where encrypted and non-encrypted Wi-Fi access points are equally distributed. The difference between the cities is still the degree of openness: New York, like Chicago, is relatively more open than other two cities, and Taipei (Fig. 11.) is the most open city.

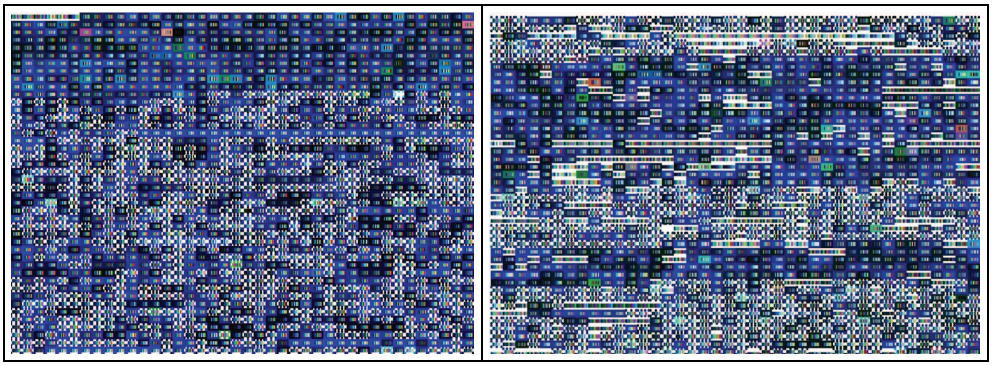

Fig. 9. Color charts of London (left) and Chicago (right).

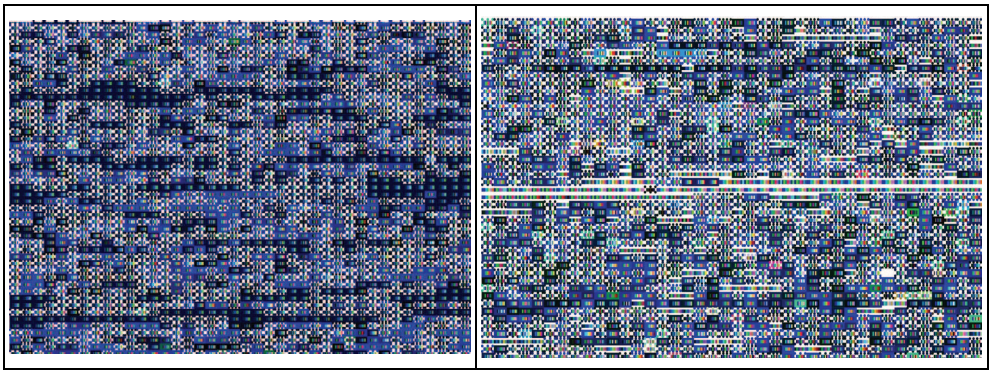

Fig. 10. Color charts of Hong Kong (left) and New York (right).

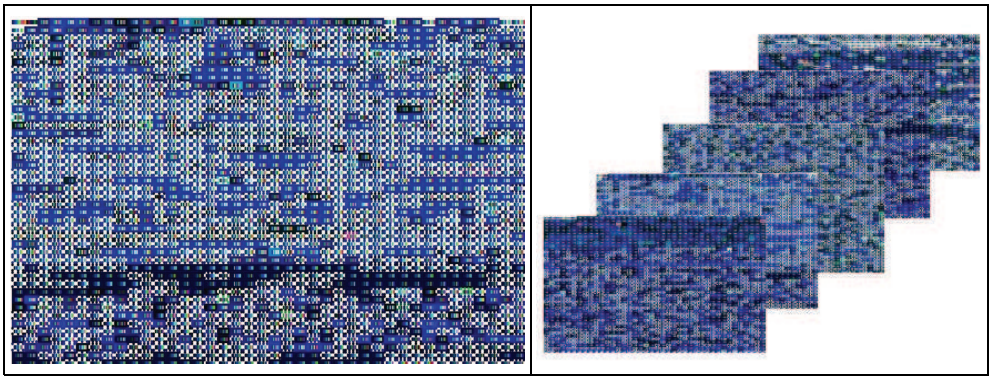

Fig. 11. Color charts of Taipei (left) and microarray inter-samples (right).

4.5 Discussion

Visualization of Wi-Fi networks presents special urban landscapes. Arjun Appadurai (1996) proposed five terms to describe the uneven and disjunctive landscape under globalization, which also applies to urban Wi-Fi landscapes. After modifying the definitions of terms in order to describe Wi-Fi networks in cities, the landscapes described below deepen the meaning of colors and patterns.

- Ethnoscapes: Wi-Fi cyborgs with similar moving routes create specific landscape of diverse cyborg group.
- Technoscapes: They can be distributions of Wi-Fi access points or personal Wi-Fi devices.
- Financescapes: The unequal Wi-Fi distribution reflects financial difference. For example, more Wi-Fi-equipped devices are affordable and needed in wealth area.
- Mediascapes: City-wide Wi-Fi infrastructure provides governments and telecom companies a new and easy way to show advertisement and messages in Wi-Fi login pages.
- Ideoscapes: Freedom, ubiquity, convenience, liberty and surveillance are the ideological landscapes constructed in Wi-Fi slogans and arguments.

By the collaboration of anthropology and linguistics knowledge, artists can communicate with other disciples with conceptualization and contextualization of artworks. The patterns and orders from aesthetics mediated Wi-Fi system to an data-mining artworks. Artworks will not be restricted to aesthetics, which implies that artists can express their idea in more elaborate and appropriate words and forms. The visualizations can also help anthropologists think pictorially to tackle complex and abstract issues with concrete and visible observations and analysis. For example, ordered and colored route charts from the Wi-Fi fieldwork data make abstract artworks conceivable and apprehensible for non-artists.

5. Conclusion

In this chapter, we discussed data mining in three disciplines: bioinformatics, anthropology and aesthetics. Data in these disciplines are usually hard to be directly visualized and hence need data mining to select and transform representative data into comprehensible forms.

Taking microarray analysis as the example in bioinformatics, Lin applied a cube mode to define inter-relations among different samples. Liu visualized the hidden Wi-Fi access points in five cities with house metaphor to analyze the similiarities and differences. Then, aesthetics concepts combined the two studies to visualize inter-sample relationships in different cities, with color charts which constitute Wi-Fi landscapes.

Visualization does not merely reflect data; rather, it offers us visual ways to re-discover important information behind complicate algorithms and huge amount of data. As artist Lev Manovich said, modern browser art "rendering the phenomena that are beyond the scale of human senses into something that is within our reach, something visible and tangible." (Manovich 2002: 8-9) Lin converted invisible gene sets to visible cubes, and therefore inter-relation of gene sets could be discovered and well-formed. Liu viewed Wi-Fi access points as houses to transform invisible Wi-Fi landscapes to visual house societies, which helped us to study Wi-Fi networks in big cities. Aesthetics made anthropology study as artworks through investigations of how patterns and rules in visualization contribute to analyze complicated phenomena. Reversely speaking, artists can also share similar language learned from data visualization with data mining experts and social scientists. Above all, researchers from various disciplines could communicate with each other and strengthen their data-mining skills through broader interpretation on visualization.

6. References

Adorno, T. W. (1997). *Aesthetic theory*, University of Minnesota Press, 081661799-6, Minneapolis

Appadurai, A. (1996). *Modernity at large: cultural dimensions of globalization*, University of Minnesota Press, 081662792-4, Minneapolis

Ahrens, K. (2002). When Love is not Digested: Underlying Reasons for Source to Target Domain Pairing in the Contemporary Theory of Metaphor. *Proceeding of the First Cognitive Linguistics Conference*, pp. 273-302, Taipei, January 2002, Cheng-Chi University, Taipei

Ahrens, K. (2010). Mapping Principles for Conceptual Metaphors, In: *Researching and Applying Metaphor in the Real World*, Cameron Lynne, Alice Deignan, Graham Low, Zazie Todd, (Ed.), 185–208 , John Benjamins,978902722380-7,Amsterdam

Ashburner, M. et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, Vol. 25, No. 1, pp. 25-29, 1061-4036

Badiou, A. (2006). Speaking the unspeakable. Original audio recording available at http://www.lacan.com/space/badiou3.mp3

Bernard, A. & Hartemink, A. (2005). Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data, *Pacific Symposium on Biocomputing*, pp. 459-470, 1793-5091, Hawaii, January 2005, World Scientific, Singapore

Borgwardt, K. et al. (2005). Protein function prediction via graph kernels. *Bioinformatics,* Vol. 32, No. 1, pp. 47-56, 1367-4803

Bourdieu, P. (1977). *Outline of a theory of practice*, Cambridge University Press, 052121178-6, Cambridge

Edgar, R. et al. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acid Research,* Vol. 30, No. 1, pp. 207-210, 0305-1048

Eisen, M. et al. (1998). Cluster analysis and display of genome-wide expression patterns, *Proceedings of the National Academy of Sciences*, Vol. 95, No. 25, pp. 14863-14868, 1091-6490

Evans-Printchard, E. E. (1940). *The nuer: a description of the modes of livelihood and political institutions of a Nilotic people*, Clarendon Press, 115352163-6, Oxford

Gollub, J. et al. (2003). The Stanford microarray database: data access and quality assessment tools. *Nucleic Acid Research,* Vol. 31, No. 1, pp. 94-96, 0305-1048

Haraway, D. J. (1991). *Simians, cyborgs, and women: the reinvention of nature*, Routledge, 041590386-6, New York

Jevbratt, L. (2002). 1:1, http://128.111.69.4/~jevbratt/1_to_1/interface_ii/index.html

Jansen, R.; Greenbaum, D. & Gerstein, M. (2002). Relating whole-genome expression data with protein-protein interactions. *Genome Research,* Vol. 12, No. 1, pp. 37-46, 1088-9051

Jones, K. & Liu, L. (2006). What where Wi: An analysis of millions of Wi-Fi access points. *Portable Information Devices, 2007. PORTABLE07. IEEE International Conference on*, pp.1-4, 142441039-8, May 2007, Orlando, FL, IEEE, Orlando

Joshi, T. et al. (2004). Genome-scale gene function prediction using multiple sources of high-throughput data in yeast saccharomyces cerevisiae. *OMICS: A Journal of Integrative Biology,* Vol. 8, No. 4, pp. 322-222, 1536-2310

Lakoff, G. & Johnson, M. (1980). Metaphors we live by, *University of Chicago Press*, 022646801-1, Chicago

Lanckriet, G. et al. (2004). Kernel-based data fusion and its application to protein function prediction in yeast, *Pacific Symposium on Biocomputing*, pp. 300-311, 1793-5091, Hawaii, January 2004, World Scientific, Singapore

Lévi-Strauss, C. (1982). *The way of the masks*, University of Washington Press, 029595929-0, Seattle

Lin, K. & Kang, J. (2007). Exploiting inter-gene information for microarray data integration. *Proceedings of the ACM Symposium on Applied Computing*, pp. 123-127, 978-1-60558-166-8, Seoul, Korea, March 2007, ACM, New York

Lin, K. et al. (2009). A cube framework for incorporating inter-gene information into biological data mining source. *International Journal of Data Mining and Bioinformatics,* Vol. 3, Issue 1, pp. 3-22, 1748-5673

Low, S. M. (1996), The Anthropology of Cities: Imagining and Theorizing the City. *Annual Review of Anthropology*, Vol. 25, pp. 383-409, 0084-6570

Malinowski, B. (1922). *Argonauts of the western Pacific; an account of native enterprise and adventure in the archipelagoes of Melanesian New Guinea*, G. Routledge & Sons, 020342126-4, London

Schmidt, T. & Townsend A. (2003). Why Wi-Fi wants to be free. *Communications of the ACM,* Vol. 46, Issue 5, pp. 47–52, 0001-0782

Tarbell, J. (2003). Substrate. Original artwork published online at http://www.complexification.net/gallery/machines/substrate

Tornow, S. (2003). Functional modules by relating protein interaction networks and gene expression. *Nucleic Acid Research,* Vol. 31, No. 21, pp. 6283-6289, 0305-1048

Torrensa, P. M. (2008). Wi-Fi geographies. *Annals of the Association of American Geographers*, Vol. 98, Issue 1, pp. 59–84, 0004-5608

Tseng, G. & Wong, W. (2005). Tight clustering: a resampling-based approach for indentifying stable and tight patterns in data. *Biometrics*, Vol. 61, No. 1, pp. 10-16, 0006-341X

Xu, L. et al. (2005). Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics,* Vol. 21, No. 20, pp. 3905-3911, 1367-4803

Zhou, X. et al. (2005). Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nature Biotechnology,* Vol. 23, No. 2, pp. 238-243, 1087-0156

Wi-Fi Alliance. (2010). Wi-Fi CERTIFIED™ Products. Homepage at http://www.wi-fi.org/certified_products.php

Knowledge-Oriented Applications in Data Mining
Edited by Prof. Kimito Funatsu

ISBN 978-953-307-154-1
Hard cover, 442 pages
Publisher InTech
Published online 21, January, 2011
Published in print edition January, 2011

The progress of data mining technology and large public popularity establish a need for a comprehensive text on the subject. The series of books entitled by 'Data Mining' address the need by presenting in-depth description of novel mining algorithms and many useful applications. In addition to understanding each section deeply, the two books present useful hints and strategies to solving problems in the following chapters. The contributing authors have highlighted many future research directions that will foster multi-disciplinary collaborations and hence will lead to significant development in the field of data mining.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Kuan-Ming Lin and Jung-Hua Liu (2011). Exploiting Inter-Sample Information and Exploring Visualization in Data Mining: from Bioinformatics to Anthropology and Aesthetics Disciplines, Knowledge-Oriented Applications in Data Mining, Prof. Kimito Funatsu (Ed.), ISBN: 978-953-307-154-1, InTech, Available from: http://www.intechopen.com/books/knowledge-oriented-applications-in-data-mining/exploiting-inter-sample-information-and-exploring-visualization-in-data-mining-from-bioinformatics-t

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2011 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License, which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.