Co-occurrence analysis as a framework for data mining

Jan W. Buzydlowski
Holy Family University

Abstract

This paper examines the use of co-occurrence analysis as the basis for, and framework of, various data mining techniques for numeric and textual data. The definition, computation, interpretation, visualization, and application of co-occurrences are discussed, as well as a survey of systems that use co-occurrence as their basis.

Keywords: data mining, analysis, co-occurrences, visualization

INTRODUCTION

Data mining is the automated exploration of data, textual or numeric, to determine rules, patterns, information, or anomalies that are unknown within a dataset. There are many methodologies that are classified as data mining, but one method that this paper will focus upon is the use of co-occurrence analysis for the extraction and display of the rules and patterns within data, both numeric and textual. More to the point, this paper will explore co-occurrence analysis as a framework for multiple data mining techniques.

Co-occurrence analysis is simply the counting of paired data within a collection unit. For example, buying shampoo and a brush at a drug store is an example of co-occurrence. Here the data is the brush and the shampoo, and the collection unit is the particular transaction. In this example, the paired data is {shampoo, brush} and it occurs once. Of course, more items can be purchased at a time, so the pairings become more numerous as each item is paired with each other item. For example, if in addition to the two items, a third item is purchased, say, goo, then there are three pairings ({shampoo, brush}, {shampoo, goo}, {brush, goo}), again each with a count of one.

The collection unit and the data therein can be varied. For example, it can be a shopping cart, with the purchased items as the data, as in the example; a questionnaire, with the responses to the questions as the elements; or a paper's bibliography, with each cited author as the focus of the analysis.

The combination of all of the collection units then forms the support for the analysis. The support may be the purchases of a particular store for a particular day, all the questionnaires returned, or a digital library of a particular journal.

For example, if 100 three-question surveys were filled out and returned as a response to an email request, then there would be 100 collection units, each containing the data for each respondent, all of them forming the support of the analysis. Again, the data is analyzed pair-wise, so if Respondent 1 answered "A" to Item 1, "B" to Item 2, and "C" to Item 3, then the data analyzed would be ({1:A, 2:B}, {1:A, 3:C}, {2:B, 3:C}). Similarly, if Respondent 2 also answered "A" to Item 1, "B" to Item 2, but "A" to Item 3, the data analyzed would be ({1:A, 2:B}, {1:A, 3:A}, {2:B, 3:A}). Both are combined and the support would be: ({1:A, 2:B}, {1:A, 3:C}, {2:B, 3:C}, ({1:A, 2:B}, {1:A, 3:A}, {2:B, 3:A}). This would be done for the remaining 98 respondents.

When items co-occur, there is an association between them by the gathering agent, e.g., a shopper, a survey participant, or an author of a journal paper citing works to support their own work. If a pairing is done only once in the support, the association is tenuous or spurious. If the pairing is done by many gathering agents, then the association is made stronger with each additional pairing.

In the previous example, for the two respondents mentioned, both responded {1:A, 2:B}. If only two answered in this way, there seems to be little association. However, if the other 98 respondents answered {{1:B, 2:A} , then there is an extremely strong relationship between Answer B for Item 1 and Answer A for Item 2. Furthering this example by supplying meaning to the questions and responses, if Item 1 was gender, where "A" meant "male" and "B" meant "female," and Item 2 corresponded to satisfaction ("A") or dissatisfaction ("B"), then, while there is a small cadre of mad males, there is a large pool of satisfied female respondents.

As another example, one which uses Author Co-citation Analysis (discussed later), if an author cites works by Plato and Woody Allen, and that co-occurrence appears only once within one paper within a large digital library, then those authors are related, but only within the mind of the one citing author. If however, Plato and Aristotle are cited thousands of times within that digital library, then it is can be inferred that those authors are highly related. It is important to note that in this example it is known to the casual reader that those associations between those two philosophers hold through prior knowledge; however, the co-occurrence analysis will show this to be true both of authors that are known, as well as unknown, i.e., mined associations. Moreover, the items need not be authors, but any two co-occurring elements, perhaps, movies that one views on a streaming video service. Some viewers have eclectic and varied tastes, so some parings will be rare, but others are more mainstream and some parings will be common and numerous. For example, few—if any—may have watched both “The Ditch Digger’s Daughter” and “Jackass Number Two,” but many have watched “Godfather” and “Godfather II.” Again, this is a theorized association, but the analysis of the data would most likely bear this out and makes sense intuitively.

Finally, the items of interest need to be determined, either *a priori*, or suggested by the data, before the analysis can be done. It would become both overwhelming and insignificant, to explore all pairings done by all entities within a collection. For example, to analyze all items purchased at a drug store by all the customers on a particular day would yield too many pairing supported by too few data. This is because the number of pairings examined grows geometrically, while the occurrence of all possible pairings is sparse. The examination of only 100 items yields 4,950 possible pairings, some of which are never purchased together by anyone, e.g., goo and hair growth tonic. (The derivation of the numbers of pairings will be discussed below). It might be of interest, then, to explore just the top-five-selling items one day, or to look at only purchases involving toothbrushes to keep the number of pairings reasonable and well supported by purchases. The last example of items associated with a toothbrush is a seed term and then use it to find other n values that co-occur with that seed; this is explored in a later section.

It is then through the examination of all the pairings of the items of interest, within a collection unit, by the gathering agents, in a support of analysis collection that patterns emerge where they then can be visualized and gleaned.

DERIVATION AND REPRESENTATION OF CO-OCCURRENCES

For the purposes of this section, a small example of specific data will be examined to show how the raw counts and/or the statistics are derived to form the basis of co-occurrence analysis. To that end, then, consider the following data:

[{A, B}, {A, B, C}, {A, D}, {B, C}]

Each set of data between the braces indicates the collection unit of analysis and the combination (four) of them all is the support of analysis, i.e., between the brackets. The items of interest will be A, B, C, and D.

The letters A though D could represent items purchased in a supermarket, so the customer, for the first set, purchased Items A and B. A second customer purchased A, B, and C. And, so on. The letters could also represent movies that four customers have viewed in the last week. The letters could also represent the key terms used to classify four papers, noun phrases within their abstracts, or cited authors (or journals) in their respective bibliography sections. It is left to the reader to generate another possible scenario which the sample data represents. What is important is to recognize that this form of data is widely available and that this paper's purpose is to show the different ways in which it can be analyzed and mined.

Once the items and collection are defined, the co-occurrence counts can then be derived. This is usually represented by a co-occurrence matrix, with the items of interest forming the row and column headings and the intersection of the row and column (cell) indicating the co-occurrence. Given our above example, then, the following matrix would be as shown in Figure 1 (Appendix).

In this representation, the values on the main diagonal are values representing the frequency of occurrence of an item within a collection. In our example, "A" occurred three times, as did "B." And so on. (Those cells allow the computation of some statistics, such as Bayesian probabilities, of the other values in a row, e.g., P (B|A) = 2/3.)

The reader will also notice that the matrix is symmetric, i.e., the value of Cell (A, B) = 2 is also the value of Cell (B, A) = 2, as co-occurrence, according to our definition, is commutative. If the interpretation of the co-occurrence, though, is given another meaning, say, temporal, i.e., {A, B} means that A was obtained first, then the matrix would not necessarily be symmetric.

Taking all of the values, the number of data points is N * N. However, many of the methods used to analyze co-occurrence matrices assume a symmetric matrix, as well as giving no meaning (significance or use) to the values of the main diagonal. Given this interpretation, suppression of the main diagonal values and a symmetric matrix, the number of unique co-occurring pairs is (N * (N - 1))/2, given N representing the number of items of interest. This, then, illustrates the computation of the example in the previous section of 4,950 pairings given 100 items of interest (= (100 * 99)/2), and how too many items under consideration soon become unwieldy, even given these restrictions of symmetry, etc. In our example there are six pair-wise data in the matrix we are considering under those two assumptions with four items of interest, as shown in Figure 2 (Appendix).

INTERPRETATION OF CO-OCCURRENCES

Once the counts are determined, then the matrix can be analyzed. It is of interest to explore the different, sometimes dual, nature of those values and how they can be interpreted.

The values so far have been interpreted as similarities, e.g., Cell (A, B) = 2 indicates that A and B are somehow related by a measure of 2, and a number of techniques use this interpretation. In this case, the higher the value, the more similar the items are. However, if the interpretation of the value is that of dissimilarity, then the matrix can be thought of as a distance matrix, similar to a driving map showing the distance between cities on a map. For example, suppose that Cell (A, B) = 2 indicates that elements A and B, say, cities, are, say, two miles, apart. Here, the larger the value the more distance the two elements are. The elements on the main diagonal then would need to be zero, as the distance between the same city necessarily needs to be zero and the matrix would be assumed symmetric. However, this paper will focus upon the values in the matrix to be that of similarity, although transformations exist for converting dissimilarity matrices to similarity matrices. This is discussed later.

Next, if instead of viewing the individual values, the rows or columns are instead focused upon, then other interpretations, and thus other methodologies applied, are also possible. For example, if the rows of a full matrix with N columns are interpreted to be a vector in N-space, then those rows represent points in that space. For example, given the full matrix example in the previous section, Point A would be (3, 2, 1, 1), Point B would be (2, 3, 2, 0), etc., in 4-Space. Given this interpretation, some techniques, e.g., self-organizing maps, a special case of neural networks, can be applied, which will be discussed later.

Another interpretation exists if the columns are defined to be responses associated with that column heading, then correlations can be calculated, say using Pearson's *rho*, between the columns. Using the calculated correlations, another matrix can be formed such that the cells represent the, say, Pearson correlation between the two rows, i.e., Cell A, B would contain the correlative measure between the column values of Row A and Row B. Cell A, C would be the correlation of Rows A and C. And so on. The values of the main diagonal would be one, because the correlation with a row with itself would be one, and the matrix would be symmetric because the correlation with, say, Rows A and C are the same as with Rows C and A. Other correlative or similarity measures also exist and can be applied, and these, along with some caveats as to their application, as well as dissimilarity measures, are discussed in (Leydesdorff, 2006).

Finally, if the values are interpreted as simply indicating a connection between the two values, ignoring the magnitude of the value, and instead replacing the value with a 1, if the original value is equal or greater than one, or 0, if the value is zero, then this becomes an adjacency matrix. For example, given our previous example, the matrix becomes, as in Figure 3 (Appendix).

Here, it is only the connection that is of interest, and the matrix can be interpreted to that representing a network, with the elements as nodes and the connections as links, where a number of techniques can be applied to analyze the data, e.g., see (Sedgewick, 1988). (Nonetheless, the value of the co-occurrences can be replaced to show the strengths of the links.) Moreover, this forms the basis of finding elements not directly related (co-occurring), but that are related through intermediate elements. Given this interpretation, social network analysis techniques, forming the basis of many new social media applications, is certainly applicable. The application of co-occurrence as a social network is discussed in (Leydesdorff, 2006). Another interpretation of indirect connections, co-occurrence chains, will be discussed later.

VISUALIZING CO-OCCURRENCE MATRICES

There are three popular ways to visually express the data of a co-occurrence matrix. They are: multi-dimensional scaling (MDS), Pathfinder networks (PFNETs), and self-organizing, or Kohonen, maps (SOMs). The format of MDS seems to be subsumed by both SOMs and PFNETs and so this paper will focus on the latter two. A reader interested in exploring MDS is referred to (Kruskal, 1978).

An n-by-n co-occurrence matrix can be viewed as both an adjacency matrix of a network with n nodes, or as a set of n vectors in n-space. As a network, a PFNET is used to render it visually (Schvaneveldt, 1990). As a vector in n-space, a SOM is used (Ritter, 1989). This section will briefly describe both.

A network consists of nodes and edges. In a co-occurrence matrix, the names themselves represent the nodes and the co-occurrence values represent the edge weights. For instance, given a simple 3 x 3 matrix in Figure 4(Appendix), a corresponding network representation is shown in Figure 5 (Appendix).

In order to keep this example simple only three nodes are used and thus contain only three edges (or links). However, in larger networks the visual complexity overwhelms the viewer and presents no advantage over the simple raw co-occurrence matrix. As mentioned, a network of 25 items (nodes) has

300 edges (values). A PFNET is used to remove some of the redundant or less-salient links to show a more understandable network.

A PFNET is created by examining each link between each node pair. Alternate paths around the two nodes are examined to see if there is a shorter path. If there is, then the link between the two nodes is eliminated. The walk length, *q*, of the alternate path (i.e., how many links to examine) and the metric used to measure the distance of the alternate path, *r*, (e.g., city-block distance, Euclidean distance, or maximum link length) are specified by the user. For this example, we use q = n-1 and r = maximum link length as this reveals a network with the fewest links.

A PFNET based on Figure 5 is shown in Figure 6 (Appendix). In this example, the link between Smith and Jones (weight of 2) is not removed as the maximum weight of the alternate path, Smith – Brown – Jones is 5, which is greater than 2. However, the link between Jones and Brown (weight of 5) is removed as the maximum weight of the alternate path, Brown – Smith – Jones, is 3, which is less than 5.

The reasoning for the removal of the link is that Jones is better related to Brown through the association of Smith rather than directly. Thus in our example the network with three links is reduced to two. In large networks, this reduction results in readily interpretable networks. The reader is referred to (Schvaneveldt, 1990) for a more thorough discussion of PFNETs and their application.

If the rows of a co-occurrence matrix are viewed as n vectors in n-space, then a special type of neural network, a self-training network, can be used to arrange, or reduce, the n-space to a lower dimension, in this case, two dimensions.

In a traditional neural network, the categories associated with an entity are known and the existing data is used to train the system to recognize new instances and place them in those categories. However, with the co-occurrence matrix it is the categories themselves which are sought and a self-organizing, or self-training, network is used to find categories by examining the data within the co-occurrence matrix.

A SOM is based on a two-dimensional grid of evenly spaced nodes. Figure 7 (Appendix) shows a network consisting of 4-by-5 nodes. (Please note that the use of network here is different from the previous discussion of PFNETs. For PFNETs, the network model is based on a mathematical definition, whereas for SOMs the network refers only to the grid of nodes.)

It is important to note that each node corresponds to a vector of values. In our example, with three names each vector would consist of three ordered values, e.g. (2, 3, 4). Each row of the co-occurrence matrix also represents a three-valued vector.

To train a SOM, the network of nodes are initialized to random values. Then, a row of the co-occurrence matrix is chosen at random and compared to each node. The node in the grid with the closest Euclidean distance to the row is determined. The values of the node chosen are modified to reduce the distance between itself and the row. Also, the nodes surrounding the chosen node are similarly adjusted. Then, another vector (row) is randomly chosen (the selection is with replacement so each vector can be reselected), and the process repeats many times.

The resulting network consists of nodes containing values corresponding to categories within the co-occurrence matrix. These categories are delineated by areas of nodes in which the element with the highest value of the vectors are the same. For example, if the second element of the vectors in the upper-right hand corner of a network are the same, then they would be considered an concept area. The area then would be labeled by the second row name, in our case, Brown, corresponding with the node with the highest of all the second value. Figure 8 (Appendix) shows this concept.

The finished networks reveal an arraignment of terms that reflect the categories inherent in the co-occurrence matrix and do so in a very concise way.

A reader interested in the further studying SOMs and their application of co-occurrence analysis is referred to (Doszkocs, 1990).

FINDING THE ITEMS OF INTEREST AND SUPPORT

As previously discussed, the more items to explore within the area of support, the more co-occurrences there are, and they grow geometrically. For the aforementioned problem involving 25 names, there are 300 co-occurrences. Moreover, there is a limit on the support that is available for each co-occurrence, as some might not co-occur at all, as well as there is a visual limit as to the number of elements that the analyst's eye can consider. For that reason, there are two major ways to determine and limit the items considered.

One method is to use an external resource to determine which items are of interest. If one was using Author Co-citation Analysis to explore the groupings of specific authors within a certain discipline, then one could research which authors were of interest and use those. Say to explore the top ten authors in bibliometrics, this would be possibly determined by their citation counts. The top ten authors within bibliometrics that are cited most often by other authors would be chosen. Or, if exploring the relationships between movies, then those that were nominated for an Academy Award for a particular year may be the items of interest. In this last case, however, it is important to note that, while the items of interest are defined, the movies, the support of analysis is not, from where the data is to be extracted. That is to say, what database will be used to find the co-occurrences of the selected movies? Where the data are to be extracted is also of importance; so, both the items to be considered as well as the database used to calculate the co-occurrence are important.

For example, every movie that is viewed by someone over the course of their NetFlix subscription can be viewed as a collection unit, and the collection of all customers within NetFlix' database would be the support of analysis. The movie rental data from Redbox could also be considered as an alternative, as well as the purchased movies within Amazon. It can be imagined that the patterns of co-occurrence may be different between these three data sources based on convenience, demographics, and purchasing/rental patterns, and the results of the analysis may be different, as well.

Finally, when people buy or consume items, their collective behavior can be considered as a collection unit. When someone is interested in a "more like this" scenario when they have seen a good movie (or read a good book), it would be easy to simply explore the database by finding people who also viewed that particular movie and then, within that set, find the second most-often (co-) occurring movie (or book). It could then be done for any number of additional movies, then, say find the 25 most frequently occurring movies, rank ordered by frequency of co-occurrence. This is the use of the aforementioned concept of finding the items of interest, that of using a seed term.

APPLICATION OF CO-OCCURRENCES

As mentioned in the beginning of this paper, data mining is the automated exploration of data, textual or numeric, to determine rules, patterns, information, or anomalies that are unknown within a dataset. A number of fields have used co-occurrence analysis to mine data that is within its purview, and three major areas will be highlighted here: bibliometrics, shopping cart analysis, and associative connections.

Bibliometrics is the study of statistical and mathematical techniques applied to the analysis of text and documents. While there are many methods to explore and analyze documents, to keep to the purpose of this paper, co-occurrence will be the primary method to explore.

The first thing to examine is the collection unit, as well as the data items. If the collection unit is the title of a paper, then the data of interest may be the nouns and noun phrases. For example, for this paper, the collection unit, the title, using nouns and noun phrases, would yield {co-occurrence-analysis, framework, data-mining}.   Another unit of collection is the key terms used to categorize a paper. For example, again for this paper, they are {data mining, analysis, co-occurrences, visualization}. A third popular unit of collection is the cited authors or cited works of a paper. Looking at this paper, one could build a collection unit of the first authors in the Bibliography Section {Balachandran, Buzydlowski, Doszkocs, etc.}. This last application is known as Author Co-Citation Analysis (ACA) and is well researched area (White, 1981) (White, 1990).

Author Co-citation Analysis looks at visualization of maps created by the analysis of the collection of co-citation patterns by authors within a particular index or database. The analyst is interested in the groupings and connections that form within the maps.

The analysis starts with the specification of the authors of interest. They could be authors of interest to the analyst, although a single name seed could be used. The next step is to find the number of times each author is co-cited with each other. A visualization is then used to show how all of the authors are related by their co-citations. What the result of such an analysis yields is the grouping of authors of interest by their areas of specialty and yields a map to show their groupings.   A specific example of an ACA, as well as methods of visualization, are discussed in later sections.

A second area of the application of co-occurrence analysis is that of shopping cart analysis.

Shopping cart analysis is the example used in the introduction to the paper. When someone buys shampoo and a brush, there is an association between the two. This is often classified as an association rule. In this form, then, the rule would be shampoo -> brush, i.e., when someone buys shampoo, then they also buy a hairbrush. The left hand side (LHS) of the rule is “shampoo.” The right hand side (RHS) of the rule is “brush.” (The converse, brush -> shampoo, may or may not be true and can be determined by metrics which will be discussed.)

The support of analysis is specified by the analyst, such as the week’s sales for a particular store. The collection unit is total sale, and the data are the items purchases. What is of interest is to find unknown rules generated by the data. For instance, in a supermarket, the rule of tomato-sauce - > spaghetti would be expected, but spaghetti -> wine may not be. If the latter was true, it may be profitable to the store to move the wine closer to the pasta.

There are statistics that are associated with association rules: support, confidence, and lift.

Support is the number of records where both elements of the rule appear (co-occur) divided by the number of records. In terms of probability, it is the probability of both elements, LHS and RHS, occurring within the support of analysis [P(LHS and RHS)], or the number of times the RHS and LHS co-occur (divided by the number of records).

Confidence is the number of records where the both left hand and right hand side of the rule appears (co-occur) divided by the number of times the left hand side of the rule appears. Again, in terms of probabilities, it is conditional probability that the RHS occurs given that the

LHS occurred [P(RHS | LHS)], or the number of times the LHS and RHS co-occur divided by the number of times the LHS occurs.

Finally, lift is the confidence of the rule divided by the number of times the support of just the RHS [P(RHS | LHS) / P(RHS)], or the number of times the LHS and RHS co-occur divided by the number of times the RHS occurs. Lift values equal or close to one indicate no association between the RHS and LHS, i.e., they are independent. Values higher than one indicate that the rule is of value. Values less than one indicate that the rule is contraindicated, i.e., the RHS occurs less often with the LHS.

The application and visualization of association rules is presented in a later section.

A third application of co-occurrence is the application of it to find indirectly related elements, i.e., exploring elements that are not directly co-occurring but are linked via intermediate co-occurring relationships.

For example, consider three cited authors: A, B, C. If A is co-cited with B, and B is co-cited with C, but C is **not** co-cited with A, then there is also an associative—perhaps unknown—link between Authors A and C through Author B (A – B – C). The chains can be longer, such as (A – B – C – D) where the not-directly connected terms do not co-occur, but are connected by terms that are, e.g., A and D do not co-occur, but A and B do, B and C do, and C and D do.

This method should also be familiar to the reader as this is the familiar “Six Degrees of Kevin Bacon” game. In this game, an actor is suggested and it is up to those involved to find the movies which involve intermediate actors which eventually connect the suggested actor to Kevin Bacon. A similar concept is to find an author’s Erdos Number, again, by finding authors that work as intermediates who co-wrote papers with the prolific mathematician, Paul Erdos. A more scholarly use of this concept is to find authors who are indirectly co-cited (Buzydlowski, 2008).

This method of indirect co-occurrences has been applied to journal keywords to determine associations previously unknown via the Arrowsmith Project (Smalheiser, 1998), where an unknown connection between an ailment and a cure was found. A system connecting co-cited authors not directly connected will be discussed in a later section.

SYSTEMS FOR CO-OCCURRENCE VISUALIZATION AND MINING

A system built to explore co-author citations was developed and was called AuthorMap (Buzydlowski, 2002). To use the system, the analyst merely entered the name of a single author of interest, a name seed, and the system responded with 24 other related names, similar to the methods described in the above section. The analyst was then able to explore the relationships between the names by using either two methods of visualization: a PFNET (Schvaneveldt, 1990) or a SOM (Ritter, 1989). The system also allowed for the retrieval of papers with the name (s) of those authors of interest for further analysis. The following figures show some of the elements that were available.

The system shown was build using a specialized data store, Noah (Buzydlowski, 2002) and was populated with data from the Arts and Humanities Index, supplied by the Institute for Scientific Information (ISI), as part of a project developed at Drexel University. The system allowed the user to enter a single name seed and retrieved 24 other related names. Figure 9 (Appendix) shows the use of “Plato” as a name seed, with the system retrieving 24 other related names.

They system then analyzed the co-occurrence of those 25 names by using two visualizations, PFNET and SOM to show how the names were related. Figure 10 (Appendix) shows the PFNET and Figure 11 (Appendix) illustrates a SOM.

Author co-citation analysis is useful for mining groups of authors within a database. From the figures one can see the groupings, such as the modern German philosophers (lower right-hand corner), the religious one (lower left-hand corner), the playwrights (upper right-hand corner), etc.

Once the authors of interest are found, it is possible within this system to retrieve the articles in which the authors are cited, as shown in Figure 12 (Appendix).

A further extension, and the third application of co-citation, of looking at the co-cited authors is to find authors that are not directly co-cited, but linked but other authors with those authors are linked to, such as the Six Degrees of Kevin Bacon example mentioned previously. A system was developed to analyze and mine those connections [Buzydlowski, 2008]. The following figures show some of the elements that are available in that system.

The system developed was command-line based, using a supercomputer, Rachel (TeraGrid, 2007), and a specialized database, Piotr (Buzydlowski, 2008). The examples shown are from (Buzydlowski, 2006) and (Buzydlowski, 2008).

The system allowed for determination of a chain linking any two names—if they were linked. So, if it was of an interest to find the linking of Sir Francis Bacon to Varro, the following chain was found:

BACON-F — CICERO — VARRO

The system also allowed for exploratory associate analysis, using a single name seed, as in AuthorMap, with the number of links specified and the support of each link also specified. For example, an interest chain based on the name seed of Emily Dickinson is:

DICKINSON-E(300)— ABBEY-E(72)— AMERY-C(31)—PEI-IM(57).

(which shows the poet linked with the architect, I. M. Pei within three links.)

Or, given a name seed of Francis Bacon, the chain:

BACON-F(896)—JONES-M(354)—AUSTRAL-BALLET(21)—SAN-FRAN-BAL(32)

joins him with the San Francisco Ballet, also within three links (The numbers in parentheses indicate the citation strength; e.g., Bacon is cited 896 times in AHCI, the San Francisco Ballet is cited 32).

While the database Piotr provides the shortest chain, it does not provide every possible chain linking the two names. Providing all of the chains would overwhelm both the user and the system.

As a compromise to every possible chain, which would contain many redundant names, the idea of orthogonal chains, chains containing unique paths (no redundant names) between the two authors of interest, would be possible.

Below is a sampled subset of the orthogonal chains generated between Emily Dickinson and I.M. Pei:

PEI-IM—AMERY-C—ABBEY-E—DICKINSON-E,
PEI-IM—ATTALI-J—ABRAHAM-N—DICKINSON-E,
PEI-IM—AULENTI-G—FOUCAULT-M—DICKINSON-E,
PEI-IM—CARRENODEMIRAND.J—JENKINS-M—DICKINSON-E.

As a final example of the visualization of co-occurrences, via association rules, is using a system MOTC (Balachandran, 1999). Using this system, the analyst would look at the distribution of the data and click on categories of interest, showing the amount of co-occurrence within other variables and groups.

A prototype for MOTC was built as a Java Applet and based on fictitious data from a fictitious university. Figure 13 (Appendix) shows the interface for eight variables.

Moving your mouse over one of the bars (a MOTC), will reveal the variable and the value. For purposes of description, each bar will be indicated by its position, MOTC row, column and the category. In Figure 13 (Appendix), assume the mouse is over the first bar and the first square (MOTC 1,1; Category 1 = Gender: Male). (The MOTC to the right would be MOTC 1,2, and the one below would be MOTC 2,1). The other seven variables are represented by a MOTC, with each category within a variable represented by a square and the area of the square indicating the number of values in the set. The variables, working left to right and down are Gender (male, female), Engage in Sex (yes, no), Age (<25, 25 – 50, 50+), Current year (freshman, sophomore, junior senior), Going to Grad School (yes, no, undecided), Do you Drink? (yes, no), Number of Siblings (1-2, 3-5, 6+), and Income Level (10,000-20,000; 20001, 30,000; 30001-40,000; 40,001-50,000; and, 50,000+). This last category shows how a continuous numeric variable can be converted to a categorical one.

Clicking on one or more of the variables shows the co-occurrence of the other related variables. For instance, clicking on Gender: Male, (MOTC 1,1; Category 1 = Gender: Male), shows the number of co-occurrence within the other variables (as indicated by a darker shading), as illustrated in Figure 14 (Appendix).

Based on the figure, one can see that all the other categories are rather uniformly distributed.

The data was constructed so as to yield three major association rules. This would be shown by looking for a larger shaded area, as compared to the other co-occurring categories. For instance, one rule is that if you are a freshman (MOTC 2, 2; first square), then you are undecided about graduate school (MOTC 3, 1; Category 3). This is illustrated Figure 15 (Appendix).

A converse to that rule is that by senior year (MOTC 2, 2; category 4), the students are decided either yes or no (MOTC 3, 1; categories 1 and 2). This is shown in Figure 16 (Appendix).

Although the example shows the interface as applied to fictitious data, a larger exposition using real-world data, as well as alternate statistics to measure hypotheses, is in (Balachandran, 1999).

CONCLUSION

When two items co-occur, there is an association between the two entities as indicated by the grouping agent. The agent can be a consumer purchasing items in a drug store, a viewer of a video streaming service, or an author of a research paper; When more than one agent also associates those two items, that is another instance which strengthens that association. The collection of all the co-occurring elements form a framework from which to mine associations, be them in the form of clusters, association rules, or transitive associations and can be found within systems which highlight for the analyst those unknown facts.

BIBLIOGRAPHY

Balachandran, K., Buzydlowski, J. , Dworman, G. , Kimbrough, S. O., Vachula, W. , Shafer, T. (1999). MOTC: An Interactive Aid for Multidimensional Hypothesis Generation. Journal of Management Information Systems. 16 (1) 17 - 36.

Buzydlowski, J. (2002). A Comparison of Self-Organizing Maps and Pathfinder Networks for the Mapping of Co-Cited Authors. Doctoral Thesis. Drexel University.

Buzydlowski, J., White, H.D., Lin, X. (2002) Term Co-occurrence Analysis as an Interface for Digital Libraries, Lecture Notes in Computer Science. (2539) 133 – 144.

Buzydlowski, Jan. (2006). Exploring Co-citation Chains, Proceedings of the American Society for Information Science and Technology, 43 (1), 1-8.

Buzydlowski, J. (2008). Six Degrees of Scholarship. Proceedings of the American Society for Information Science and Technology. 45 (1) 1-4

Doszkocs, T.E., Reggia, J., Lin, X. (1990). Connectionist Models and Information Retrieval. Annual Review of Information Science and Technology. 25, 209-260.

Kruskal, J.B., Wish, M. (1978). Multidimensional Scaling. Sage.

Leydesdorff, L., & Vaughan, L. (2006). Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment. Journal of the American Society for Information Science and Technology. 57(12), 1616-1628.

Ritter, H., Kohonen. T. (1989). Self-Organizing Semantic Maps. Biological Cybernetics 61, 214-254.

Schvaneveldt, Roger W., ed. (1990). Pathfinder Associative Networks. Ablex.

Sedgewick, Robert, (1998) Algorithms. 2nd Edition, Addison-Wesley Publishing.

Smalheiser, N.R., Swanson, D.R. (1998). Using ARROWSMITH: A Computer-assisted Approach to Formulating and Assessing Scientific Hypotheses. *Computer Methods and Programs in Biomedicine.* 57(3), 149-153.

TeraGrid: Analysis of Organization, System Architecture, and Middleware Enabling New Types of Applications, C. Catlett et al., HPC and Grids in Action. (2007) L. Grandinetti, ed., Advances in Parallel Computing Series, IOS Press, Amsterdam.

White, H. D., Griffith, B. C. (1981). Author Cocitation: A Literature Measure of Intellectual Structure. Journal of the American Society for Information Science. 32, 163-172.

White, H. D. (1990). Author Cocitation Analysis: Overview and Defense. In Bibliometrics and Scholarly Communication, Christine Borgman, ed. Newbury Park, CA: Sage. 84-106.

APPENDIX

Figure 1

	A	B	C	D
A	3	2	1	1
B	2	3	2	0
C	1	2	2	0
D	1	0	0	1

Figure 2

	A	B	C	D
A	---	2	1	1
B		---	2	0
C			---	0
D				---

Figure 3

	A	B	C	D
A	---	1	1	1
B		---	1	0
C			---	0
D				---

Figure 4

	Smith	Brown	Jones
Smith	10	3	2
Brown	3	23	5
Jones	2	5	9

Figure 5

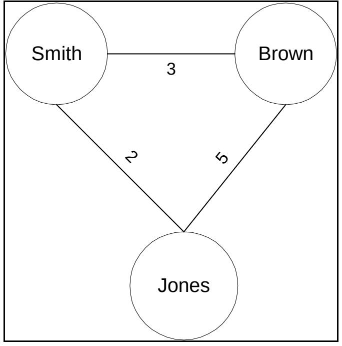


Figure 6

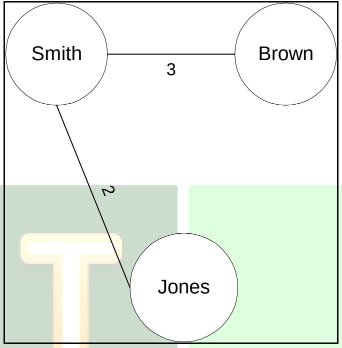


Figure 7

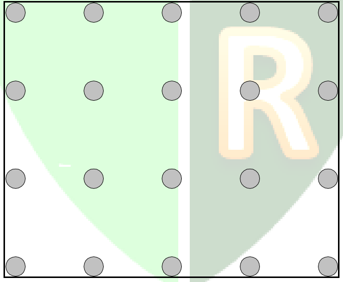


Figure 8

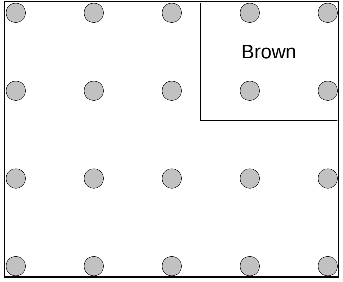
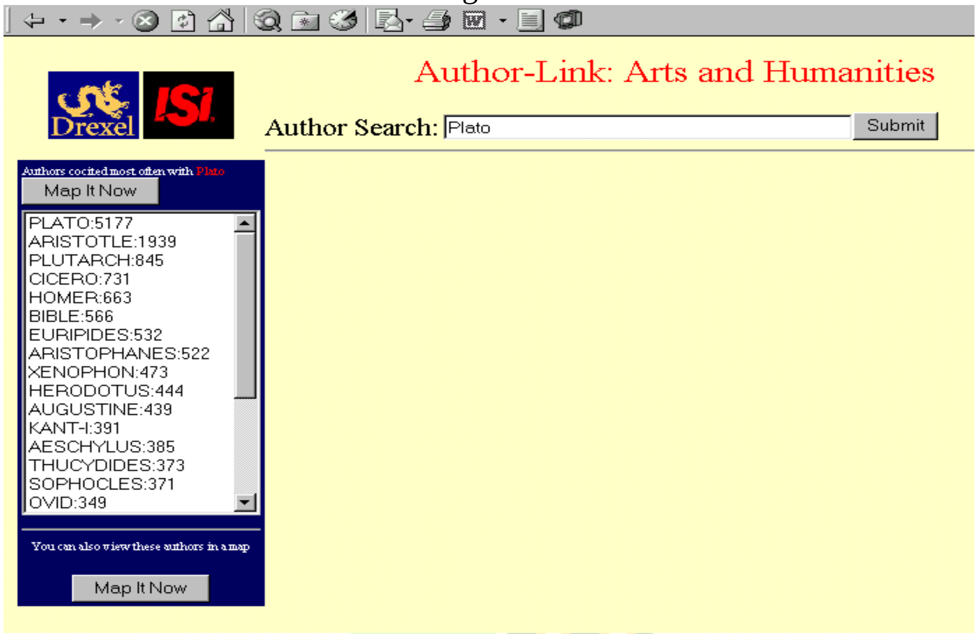


Figure 9



Author-Link: Arts and Humanities

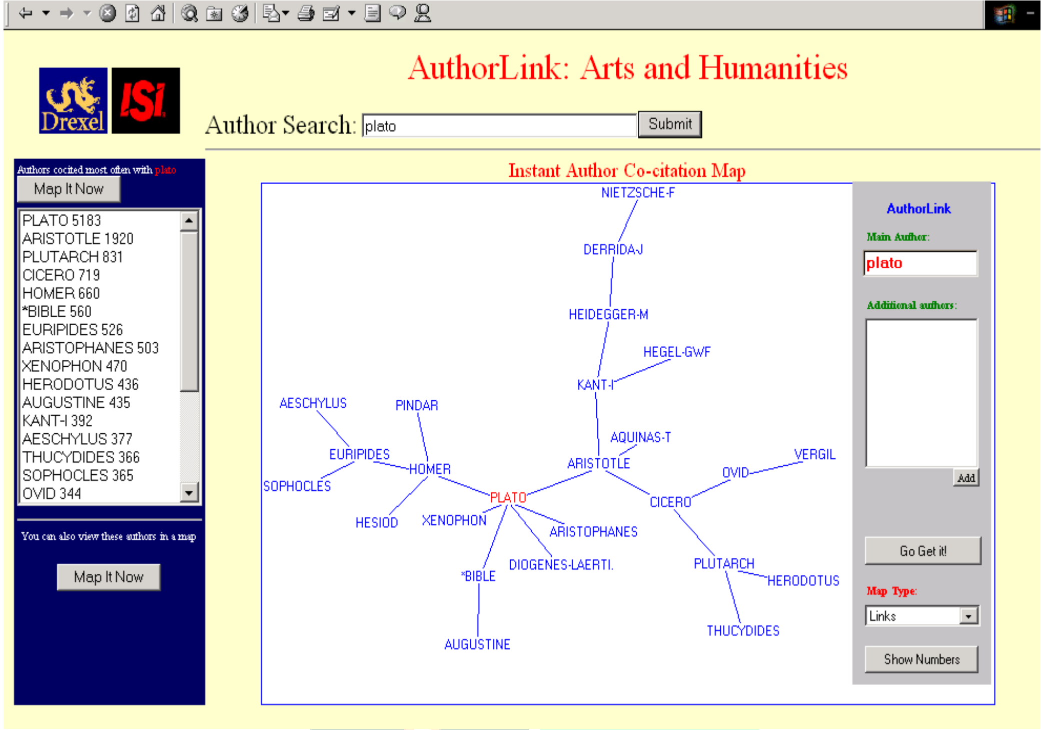
Author Search: Plato

Authors cocited most often with Plato

PLATO:5177
ARISTOTLE:1939
PLUTARCH:845
CICERO:731
HOMER:663
BIBLE:566
EURIPIDES:532
ARISTOPHANES:522
XENOPHON:473
HERODOTUS:444
AUGUSTINE:439
KANT-I:391
AESCHYLUS:385
THUCYDIDES:373
SOPHOCLES:371
OVID:349

You can also view these authors in a map

Figure 10



AuthorLink: Arts and Humanities

Author Search: plato Submit

Authors cocited most often with plato

Map It Now

PLATO	5183
ARISTOTLE	1920
PLUTARCH	831
CICERO	719
HOMER	660
*BIBLE	560
EURIPIDES	526
ARISTOPHANES	503
XENOPHON	470
HERODOTUS	436
AUGUSTINE	435
KANT-I	392
AESCHYLUS	377
THUCYDIDES	366
SOPHOCLES	365
OVID	344

You can also view these authors in a map

Map It Now

Instant Author Co-citation Map

AuthorLink

Main Author: plato

Additional authors:

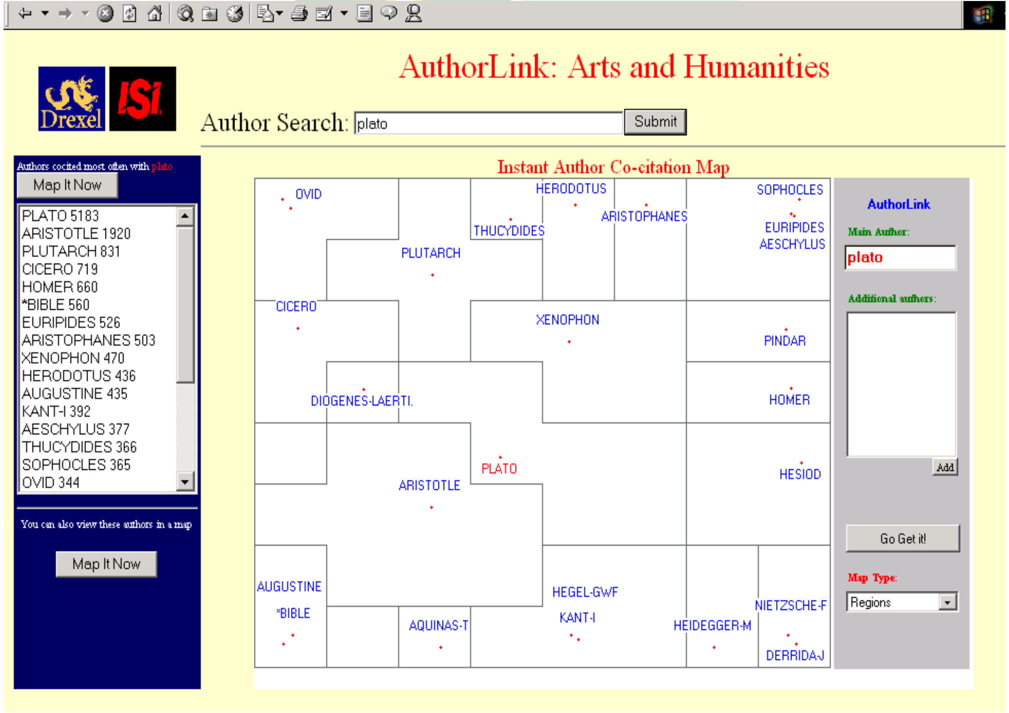
Add

Go Get it!

Map Type: Links

Show Numbers

Figure 11



AuthorLink: Arts and Humanities

Author Search: plato Submit

Authors cocited most often with plato

Map It Now

PLATO	5183
ARISTOTLE	1920
PLUTARCH	831
CICERO	719
HOMER	660
*BIBLE	560
EURIPIDES	526
ARISTOPHANES	503
XENOPHON	470
HERODOTUS	436
AUGUSTINE	435
KANT-I	392
AESCHYLUS	377
THUCYDIDES	366
SOPHOCLES	365
OVID	344

You can also view these authors in a map

Map It Now

Instant Author Co-citation Map

AuthorLink

Main Author: plato

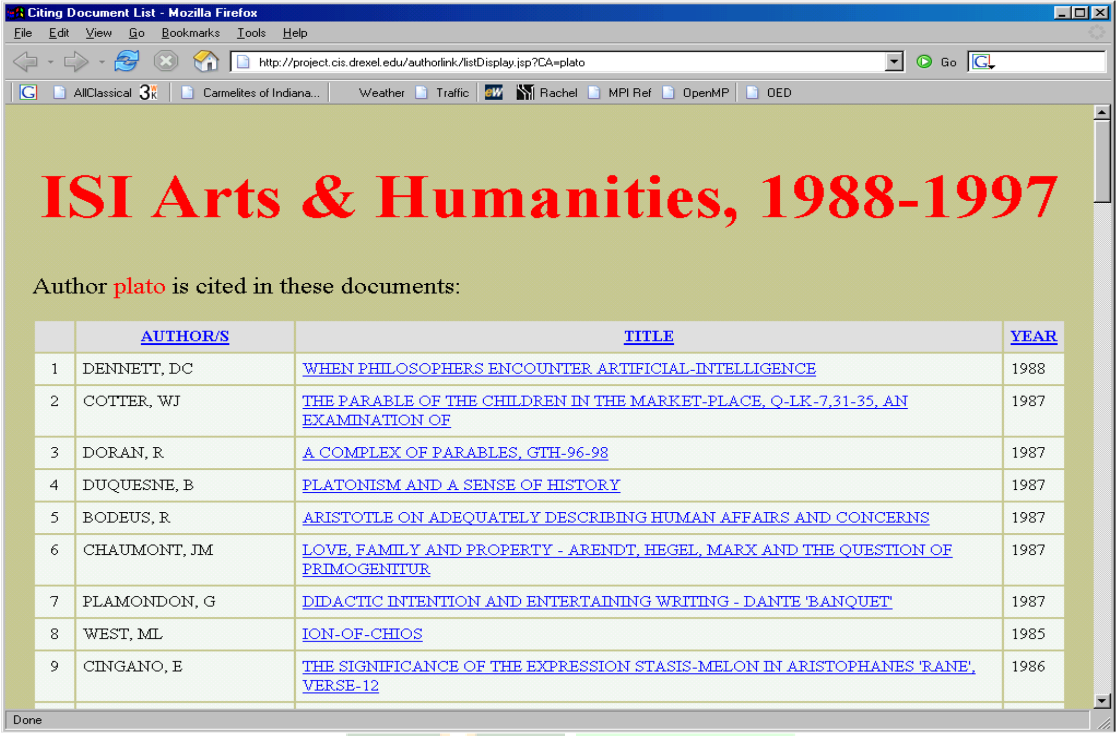
Additional authors:

Add

Go Get it!

Map Type: Regions

Figure 12



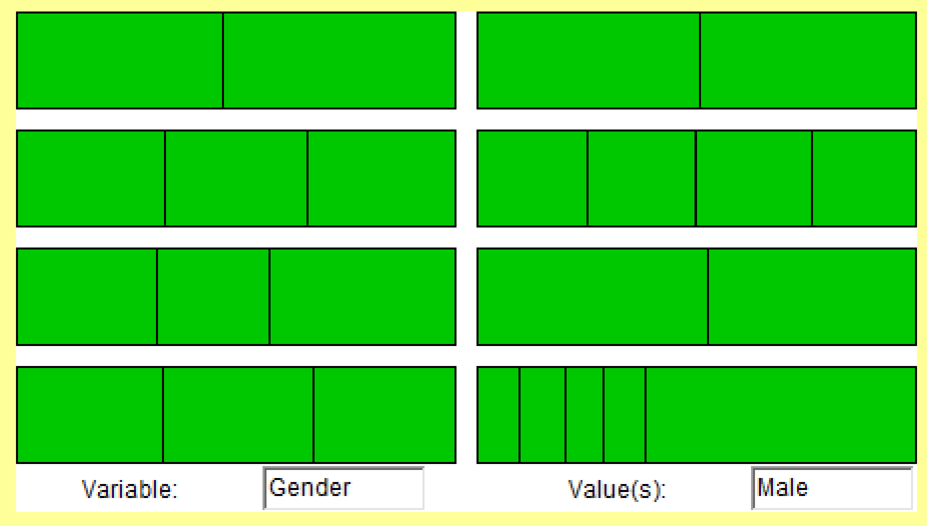
Citing Document List - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://project.cis.drexel.edu/authorlink/listDisplay.jsp?CA=plato
Go
AllClassical
Carmelites of Indiana...
Weather
Traffic
Rachel
MPI Ref
OpenMP
OED
Done

ISI Arts & Humanities, 1988-1997

Author plato is cited in these documents:

	AUTHOR/S	TITLE	YEAR
1	DENNETT, DC	WHEN PHILOSOPHERS ENCOUNTER ARTIFICIAL-INTELLIGENCE	1988
2	COTTER, WJ	THE PARABLE OF THE CHILDREN IN THE MARKET-PLACE, Q-LK-7,31-35, AN EXAMINATION OF	1987
3	DORAN, R	A COMPLEX OF PARABLES, GTH-96-98	1987
4	DUQUESNE, B	PLATONISM AND A SENSE OF HISTORY	1987
5	BODEUS, R	ARISTOTLE ON ADEQUATELY DESCRIBING HUMAN AFFAIRS AND CONCERNS	1987
6	CHAUMONT, JM	LOVE, FAMILY AND PROPERTY - ARENDT, HEGEL, MARX AND THE QUESTION OF PRIMOGENITUR	1987
7	PLAMONDON, G	DIDACTIC INTENTION AND ENTERTAINING WRITING - DANTE 'BANQUET'	1987
8	WEST, ML	ION-OF-CHIOS	1985
9	CINGANO, E	THE SIGNIFICANCE OF THE EXPRESSION STASIS-MELON IN ARISTOPHANES 'RANE', VERSE-12	1986

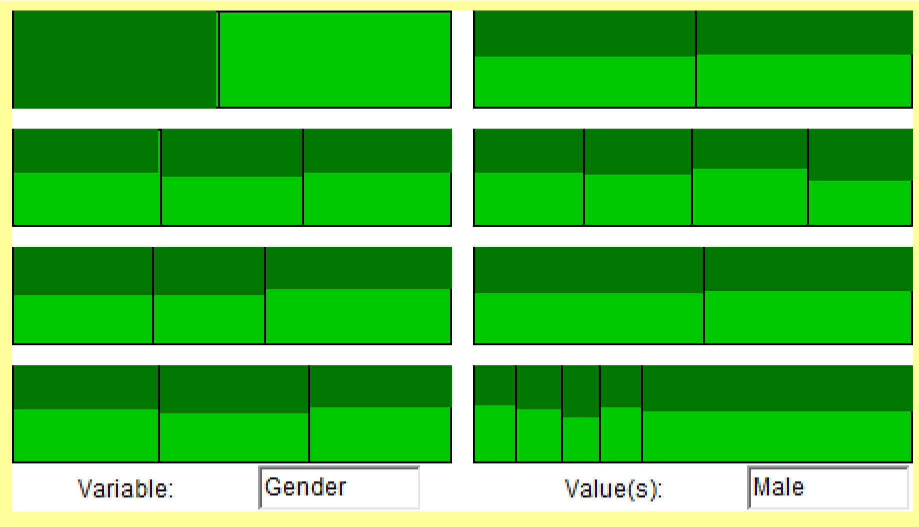
Figure 13



Variable: Gender

Value(s): Male

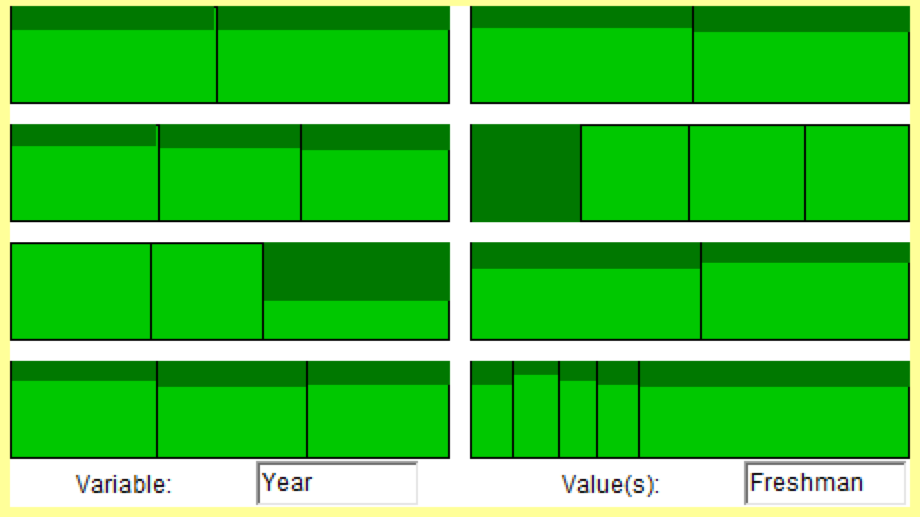
Figure 14



Variable: Gender

Value(s): Male

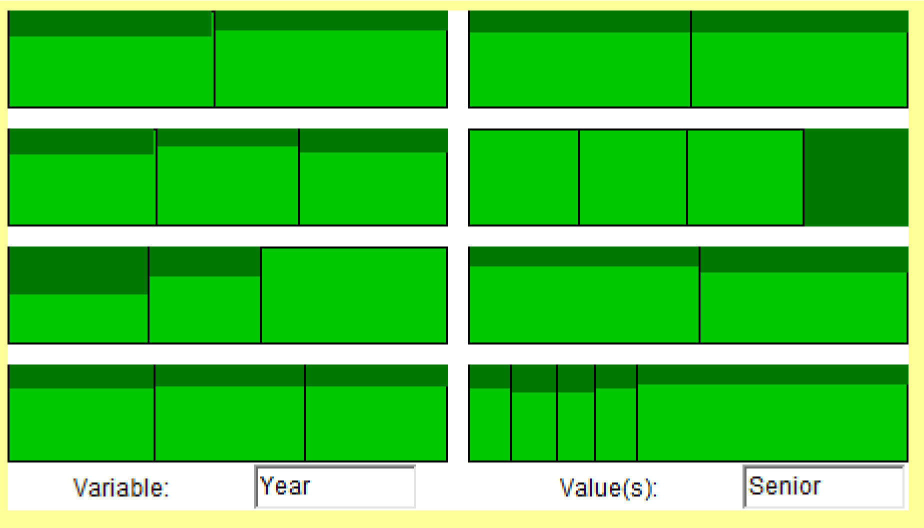
Figure 15



Variable: Year

Value(s): Freshman

Figure 16



Variable: Year

Value(s): Senior