Co-Word Analysis Tool

Ivan Krsul (ivan@krsul.org)

1   Co-Word Analysis

Co-word analysis is a content analysis technique that is effective in mapping the strength of association between keywords in textual data. Co-word analysis reduces a space of descriptors (or keywords) to a set of network graphs that effectively illustrate the strongest associations between descriptors [29, 30].

Co-word analysis is an example of a graphical modeling technique that applies some of the ideas of association analysis [27, 28]. Graphical modeling is a variant of statistical modeling that uses graphs to display models. "In contrast to most other types of statistical graphics, the graphs do not display *data,* but rather and interpretation of the data, in the form of a *model*… Graphs have long been used informally… to visualize relations between variables." [27].

This technique illustrates associations between keywords by constructing multiple networks that highlight associations between keywords, and where associations between networks are possible.

2   Co-Word Analysis Algorithm Description

In this section we describe the algorithms used in [29] for constructing the networks that highlight the strongest associations between keywords, modified to fit our needs.

Two keywords co-occur if they are used in the same records in a database. Let c_k represent the number of occurrences of the keyword k in the entire database. Let c_{ij} represent the number of co-occurrences of keywords i and j. The strength S of association between descriptors i and j is given by the following equation:

S: **integer** × **integer** × **integer** → **real**
fun $S(c_{ij}, c_i, c_j) :=$
      if $((c_i \leq 0) \vee (c_j \leq 0))$ **then**
          $S := undefined;$
      else
          $\Rightarrow S \text{ is in the range } 0 \leq S \leq 1 \Leftarrow$ *Equation 4*

$$S := \frac{c_{ij}^2}{c_i c_j}$$

      fi
    nuf

Keywords that often appear together will have strengths closer to 1, and keywords that appear together infrequently will have strengths closer to 0. In co-word analysis, strengths of larger value constitute the links between nodes in a network depicting the strongest associations in the database.

The co-word algorithm uses two passes through the data to produce the desired networks. The first pass constructs the networks depicting the strongest associations, and links added in this pass are called internal links. The second pass adds to these networks links of weaker strengths that form associations between networks. The links added during the second pass are called external links.

Note that two keywords that appear infrequently in the database but always appear together will have larger strength values than keywords that appear many times in the database almost always together. Hence, possibly irrelevant or weak associations may dominate the network. A solution to this problem –incorporated into the algorithm described in this section– is to require that only the keyword pairs that exceed a minimum co-occurrence are considered potential links while building networks during the first pass of the algorithm.

During the first pass, the link that has the largest strength is selected first, its nodes becoming the starting nodes of the first pass-1 network. Other links and their corresponding nodes are added to the graph using a breadth-first search on the strength of the links (i.e. the strongest link connecting a node that is not in any graph to the graph being constructed is added first), until there are no more links that exceed the co-

occurrence threshold, or a maximum pass-1 link limit is exceeded. The next network is generated in a similar manner starting with the link with the largest strength that is not in any existing graph.

During the second pass of the algorithm we add nodes to each existing graph, choosing the links that have the largest strength that exceed the co-occurrence threshold, and that are in some pass-1 network.

As described in [29], the algorithm for the generation of the networks is as follows:

1. Select a minimum for the number of co-occurrences, c_{ij}, for descriptors i and j, select maxima for the number of pass-1 links, and select maxima for the total (pass-1 and pass-2) links;

2. Start pass-1;

3. Generate the highest S value from all possible descriptors to begin a pass-1 network;

4. From that link, form other links in a breadth-first manner until no more links are possible due to the co-occurrence minima or to pass-1 link or node maxima. Remove all incorporated descriptors from the list of subsequent available pass-1 descriptors;

5. Repeat steps 3 and 4 until all pass-1 networks are formed; i.e., until no two remaining descriptor pairs co-occur frequently enough to begin a network;

6. Start pass-2;

7. Restore all pass-1 descriptors to the list of available descriptors;

8. Start with the first pass-1 network.

9. Generate all links to pass-1 nodes in the current network to any pass-1 nodes having at least the minimal co-occurrences in descending order of S value; stop

when no remaining descriptor pairs meet the co-occurrence minima, or when the total link maxima is met. Do not remove any descriptors from the available list;

10. Select the next succeeding pass-1 network, and repeat step 9.

Networks are interconnected by pass-2 links. The **centrality** of a network measures the degree of interaction to other networks and is defined as the square root of the sum of the squares of the S values of the pass-2 links of the network. The **density** of a network measures the internal strength of the network and is defined as the mean of the S values of the pass-1 links of the network.

Isolated Networks are those that have low centrality values. **Principal Networks** are those that have high centrality and high density values.

3   Selection of Keywords

For co-word analysis, every field in a database can be used as a keyword, or can be transformed to a series of keywords by applying the following rules:

1. If a field in the database can have the values yes, no, ?, and NA[1], then a keyword with the name of the field is generated if the value of the field is yes.

2. If the field in the database can have a single value from a list then a keyword with the name of the field followed by the value of the field is generated.

3. If the field in the database can have a list of values from a well-defined set, then for every value in the field we generate a keyword with the name of the field followed by the value of the field. G

4   The Results of the Co-Word Tool

The co-word tool was originally developed for [24] at the COAST Laboratory at Purdue University and was ported to the Windows 95/98/2000 operating systems. Figure 2 illustrates a graph produced by the tool.
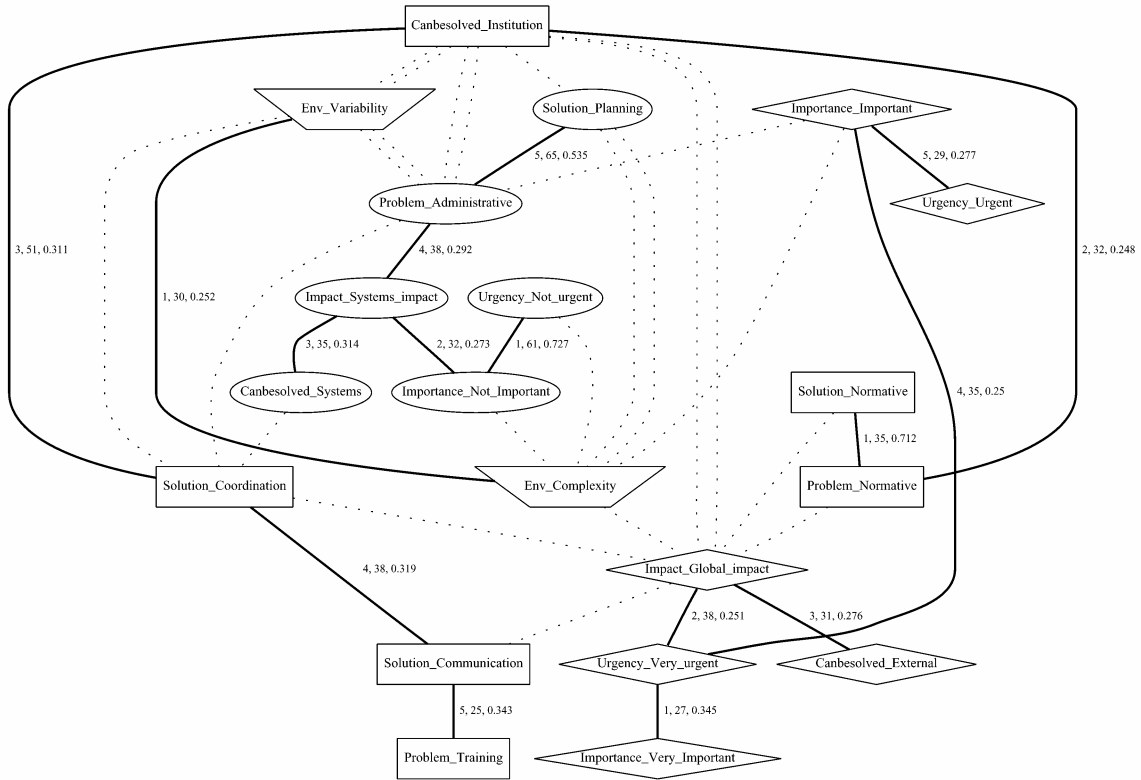
---

[1] NA: Does not apply.

Figure 2: Interconnected networks for max pass 1 links = 5, max links = 10, min coword = 25

The networks generated by the co-word analysis tool show the distribution of their centrality and density values.

In these networks, solid lines connecting nodes represent pass-1 links and dotted lines represent pass-2 nodes. Each link is labeled with a triple *<number, co-occurrence, strength>* that indicates the order in which the links were added (1 corresponds to the first link), the co-occurrence for the key pair, and the strength of the link.

The resulting networks that have the highest centrality and density values represent the more predominant relationships among the keywords in the data.

Isolated networks identify relationships with keywords that have low centrality values and hence the keywords are infrequently used in other networks.

Isolated networks with high-density values indicate strong relationships among keywords in isolated groups, and point to dominant features of the database.

Bibliography

[24] Ivan V. Krsul, *Software Vulnerability Analysis*, Ph.D. Thesis, Purdue University, May 1998.

[27] D. Edwards, *Recent Advances in Descriptive Multivariate Analysis*, Royal Statistical Society Lecture Note Series, Chapter 7, *Graphical Modelling*, pages 135-156, Clarendon Press, 1995.

[28] Leonard Kaufman and Peter J. Rousseeuw, *Finding Groups in Data*, John Wiley & Sons, Inc., Wiley Series in Probability and Mathematical Statistics, 1990.

[29] Neal Coulter, Ira Monarch and Suresh Konda, *Software Engineering as Seen Through Its Research Literature: A Study in Co-Word Analysis*, Journal of the American Society for Information Science (JASIS), Volume 49, Number 13, Pages 1206-1223, November 1998.

[30] J. Whittaker, *Creativity and Conformity in Science: Titles, Keywords, and Co-Word Analysis*, Social Science in Science, Volume 19, Pages 473-496, 1989.