Interesting Instance Discovery in Multi-Relational Data

Shou-de Lin
Information Sciences Institute, University of Southern California
4676 Admiralty Way, Suite 1001,Marina Del Rey, CA, 90292
sdlin@isi.edu
Advisor: Dr. Hans Chalupsky

The general area of machine discovery focuses on methods to use computers to perform or assist discovery tasks. Herbert Simon described it as "gradual problem-solving processes of searching large problem spaces for **incompletely defined goal objects**" [Simon, 1995, p.171]. Today machine discovery research falls into two major categories, scientific discovery and knowledge discovery and data mining (KDD). In this paper we propose a new research direction that lies somewhere in-between these two trends: we call it **interesting instance discovery (IID)** which aims at discovering interesting instances in large, multi-relational datasets.

There are three important characteristics for IID research: (1) Unlike scientific discovery and KDD, it aims at the discovery of particular interesting *instances* as opposed to general laws or patterns. (2) It is dealing with multi-relational data instead of numeric data that is best described as a relational graph or a semantic net. In such a network nodes represent objects and links represent relationships between them – see Figure 1 for an example of such a network from a bibliography domain. (3) Similar to KDD, it also focuses on data that are too large and complex to be analyzed manually by humans.

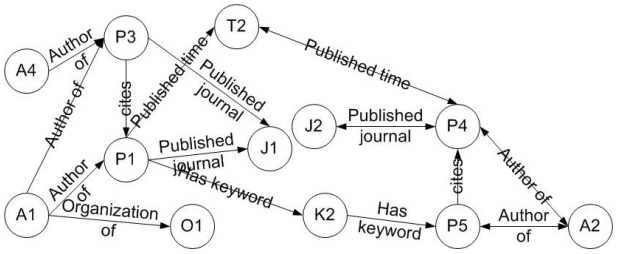

Figure 1: Example multi-relational dataset from a bibliography domain represented as a graph

The main challenge of IID arises from the fact that the term "interestingness" is vague so there is no consensus on how it can be measured. Therefore it is hard to find unbiased training examples for learning. This incompletely defined goal object makes IID a discovery problem instead of an easier supervised learning problem. In addition, the lack of universally accepted interestingness measures also creates a difficult problem on how to evaluate and verify the results of an IID program.

Our primary research goals try to address these challenges. They are as follows:

1. Investigate unbiased, universal features that (at least partially) can capture the "meaning" or essential characteristics of instances (i.e. the nodes and links in the network as shown in Figure 1) in multi-relational data.
2. Research how such features can be exploited to identify interesting instances.
3. Explore methodologies to verify IID results.

Potential Applications for IID

We believe that interesting instances play an important role in many areas such as fraud detection, intrusion detection, criminal investigation or homeland security. Since illegal and covert activities are generally less frequent and more unusual than normal activities, we believe that they are more likely to generate "interesting" evidence. Therefore, an interesting entity or event might be a good indicator for a dangerous threat, a credit card fraud, an unauthorized computer intrusion, etc. To identify those special entities, we need a methodology that aims at identifying interesting *instances* instead of patterns. This is particularly important, since in situations such as the ones described above, we will often need to find things without knowing exactly what to look for.

More generally, we believe that interesting instances could serve as "inspirations" in the inspiration-driven discovery process depicted in Figure 2. In this process people first have some problems in mind but have no clear idea what a solution could be. Suddenly they notice something interesting (we call it inspiration) that triggers the formulation of some hypothesis (or potential solution). One then has to usually look for verification to either prove or disprove the hypothesis. Inspirations play an important role in this discovery process and history shows that interesting instances have often served as the inspiration for discoveries. For example, the various types of bird beaks Darwin saw on Galapagos triggered his idea of natural selection.

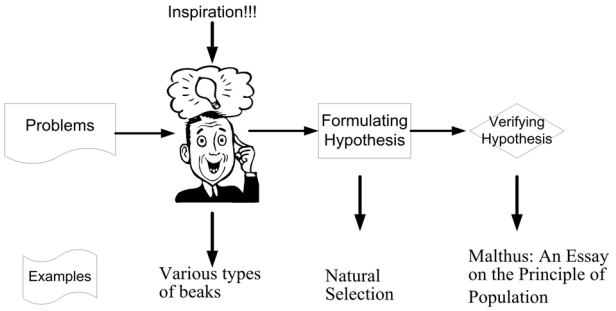

Figure 2: The process of inspiration-driven discovery

Potential Solutions for IID

So far we have developed two unsupervised methods to address our first and second research goals. To find interesting instances without training examples, we need to think about what could be the general characteristics that make one instance more interesting than another. Here are two potential solutions:
(1) A node is interesting if it carries different, abnormal semantics compared to others [Lin & Chalupsky, 2003b].
(2) A path (or loop) is interesting if it is rare, or, if it occurs extremely frequently [Lin & Chalupsky, 2003a].

The challenge for (1) is how to capture the semantics of a node as well as to identify the ones with abnormal semantics. We observe that each path surrounding or emanating from a node carries certain semantics of the node and we quantify that by computing the statistical "contribution" of each type of path for a given node participating in it. We then find abnormal nodes by looking for outliers in the contribution domain. The basic idea is that a node is abnormal if it contributes differently than others with respect to the types of paths in which it participates.

The challenge for (2) is that in fact each path occurs exactly once in the network, thus they are all equally rare. We therefore define the rarity of a path as the reciprocal of "similar" paths in the network, and our program provides the users four different choices for path similarity.

In general we tried to tackle this problem by integrating insights from both symbolic and statistical AI. The logic representation for an instance is applied to generate feature sets and we then compute their influence statistically to use as feature values.

Verification of IID Results

The lack of training examples and the incompletely defined goal object are two intrinsic characteristics that distinguish a discovery problem from a learning problem. These make a discovery problem not only more difficult to carry out but also to evaluate and verify its results. In IID we encounter a chicken and egg dilemma, since if there were unbiased ways to judge whether an instance is interesting or not, one could simply implement them to search for the solution. Lack of such universally agreed upon measures, however, prevents us from verifying discovered results directly. Since evaluation of IID results is still important, we want to investigate several indirect ways that could strengthen our belief in the validity of our IID system:

1. Rediscovery: The spirit of rediscovery lies in the verification of the methodology itself. Since there is no gold standard for verifying the results, one detour we can take is to check whether the methodology itself is applicable to a similar domain or data for which we have a better understanding. We have applied this method in the past by showing that our IID method could discover pre-defined events of interest (e.g., gang wars) in a synthetic dataset about organized crime.
2. Explanation-based discovery: The idea of this approach is to develop discovery systems that not only produce the discovered results but also generate explanation (in natural language or other human-understandable form) describing how and why the program discovered the results.
3. Minimum description length (MDL): MDL is widely applied to guide learning methods and might have similar uses for discovery. The basic idea behind it is that we prefer to store information in as little capacity as possible. The MDL criteria for verifying discovered pattern prefers simple patterns that cover the majority of data. We are investigating whether this idea can also be applied to verify IID results.
4. Exploiting independent sources: Knowledge can be represented in various forms as well as acquired from difference sources. Take the sentence "a bachelor is male" for example: one could prove it true by logical reasoning based on a certain existing ontology. Or one could say it is correct, since of a 1000 people asked 99% agreed with it. Or one could claim it is correct because there are 9 relevant documents generated by Google when using "bachelor is male" as a keyword. This shows that logical inference, statistic evidences or the Web could all serve as different ways to verify a particular piece of knowledge. Similarly, to verify a discovered result, one can try to explain the results from these views whichever are independent of the discovery methods. For example, in [Lin & Chalupsky, 2003b] we tried to verify our results on the KDD Cup 2003 bibliography dataset by (1) manually checking the semantics (note that our IID program does not understand the semantics of those relationships as human beings do) and (2) by trying to find supporting evidence on the Web.

Summary

We proposed a new discovery problem called interesting instance discovery, which has applications in many areas such as homeland security or scientific discovery. Difficult challenges are what it means for an instance to be interesting as well as how to verify discovered results, but initial results show promising success [Lin & Chalupsky, 2003b]. Other issues we want to address are automatic explanation of results, how to handle temporal information and noise as well as scalability of our algorithms.

References

[Lin & Chalupsky, 2003a] S. Lin and H. Chalupsky. "Unsupervised Link Discovery in Multi-relational Data via Rarity Analysis". In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM '03)*. 2003.

[Lin & Chalupsky, 2003b] S. Lin and H. Chalupsky "Using Unsupervised Link Discovery Methods to Find Interesting Facts and Connections in a Bibliography Dataset". in *SIGKDD Explorations*, 5(2): p.173-178. December 2003.

[Simon, 1995] H. Simon. "Machine Discovery". *Foundations of Science*, 1(2), p.171-200. 1995.