believed to have the greated potential for the discovery of knowledge, in such areas as biology and biomedicine, as it addresses parameters which are forever present and inseperable, but do change. It is noted that Bacon and Mill were dealing with complete data, while GUHA, Blum and others, deal with this incomplete data. It is considered that just because missing or noisy data was not an issue, that this is no reason that the methods should not be adapted to seek out aspects of “interestingness” in complex databases. Plotkin’s assertion of the need for inspiration is recognised by Bacon and Mill and all imply a need to develop extensible and interactive systems.

Blum

Blum, in the R X Project (1982), proposed an implementation to automate the process of hypothesis generation and the exploratory analysis of data in large databases. It was explicitly for the discovery and representation of causal relationships in clinical databases. The use of the system required detailed knowledge of clinical medicine, experimental design and statistics. The fact that “facts” are incorporated into the knowledge base, on the base of previous passes establishes it as forfilling the method of residues. Although it is not clearly stated, the RX Project is an interactive system.

Plotkin

Plotkin in 1971 undertook to develop an algorithm to find the least generalisation of clauses, into a theory of inductive generalisation. He disregards Bacon and Mill by stating that they are “an unattainable dream”, and then proceeded to state that his work “is meant to be a study of the weakest possible kind of hypotheses formation”. He further states that the essential problem is that of moving from the particular to the general. Bacon on this subject was well aware of the problem, and stated that “to conclude upon an enumeration of particulars without instances contradictory is no conclusion, but a conjecture; for who can assure .... that there are not others on the contrary side which appear not.” Further, Plotkin asserts that any reasonably successful method of inductive generalisation requires inspiration.

5. Filling the gaps

Where it is clear that, broadly, the stated methods aimed to determine new information or hypotheses from databases, it is also clear that their approaches are markedly different.

We find that if Bacon and Mill’s expositions are an accurate account of the task of elicitation of new knowledge from complex data, then it will inevitably follow that all methods to determine new knowledge, if they are to be successful, will have to have, at least in part, something in common with the expositions of Bacon and Mill. The one

and Data Engineering, on the specific matter of Learning and Discovery in Databases. While a consideration of all of these approaches would be of interest, space does not permit.

It should be noted that there are approaches which have been proposed earlier and suggested outside of this arena which we believe have a significant bearing on the subject. Given that it is the case that Bacon and Mill are generally correct in their formulation of the isolation of causal connections for discovery then, inevitably consciously or otherwise, all approaches will have important features in common with Bacon and Mill's expositions. The question which thus arises is, that of completeness. Do the methods as proposed, fully incorporate Mill's Methods? In considering this question, we turn to two methods which have features, most clearly in common with Mill's Methods and consider some of the criticisms of Plotkin.

General Unary Hypotheses Automaton

The General Unary Hypotheses Automaton (GUHA) method was presented in 1966 as an application of 'mathematical' logic to research problems of concrete sciences. It was proposed as an application to all problems where it is required to obtain unknown laws, relations or causal connections. The basis of mathematical logic make it possible to describe all assertions which might be hypotheses. The aim was to determine generally valid relations such as dependencies and connections between properties. The method was proposed to came in at stage three of a 'research' triangle, where the result is that the researcher is provided with a listing of all assertions which are valid in the model. One applies these to the entire dataset, to check their veracity to determine the necessity and direction of further research. The function of the GUHA is to provide hypotheses not verify them, yet it was designed specifically not to be interactive. Generally it can be stated that this approach broadly achieves the criteria of the methods of agreement, indirect difference and to an extent, the method of residues. The GUHA approach, filters what is explained on the basis of what is discovered during the current execution.

Method of Residues

This method is essentially based upon the consideration of experience. That is to say, where it is known from previous investigations that some of the circumstances present are known to result in such other circumstances being present, then these can in effect be discounted from the investigation, to concentrate ones attention to a smaller collection of circumstances. The known information can be acknowledged and the remaining circumstances and phenomenon can be linked in a causal connectivity relationship. It is interesting in that it calls upon antecedently established causal laws and that it examines or only needs to examine one instance, this not being the case for the other methods. Accordingly, this approach can only rightly be used to suggest probable cause. It should also be noted that it does depend on the veracity of the antecedently established causal laws.

Method of Concomitant Variation

What this method addresses, which the others do not, is that one can not always eliminate a phenomenon or circumstance from an investigation. We therefore consider the intensity of an effect, as Mill extrapolated in his example of the moon causing the variation in the tides. This method admits more evidence accordingly widening the range of inductive inference. It is the quantitative method of inductive inference, while the others are qualitative.

This method is interesting in that it is not necessarily required that an increase in ‘A’ will cause an increase in ‘a’, it may be the case that an increase in ‘A’ results in a decrease in ‘a’. It should be noted that the method implies longitudinal paired samples and that independent samples were not thought of or considered. It is further asserted that the method implies ordinal variations.

2) Depended on experience, of a restricted and trivial sort.

Mill's Methods

In 1843, John Stuart Mill (1806 - 1873), in his System of Logic, formulated the exposition of Sir Francis Bacon (1561 - 1626) as stated in his Novum Organum (1620), and it is this now classic formulation which is termed "Mill's Methods" of inductive inference. Mill defined induction as the operation of discovering and proving general propositions.

Mill's Formulation

There are four methods of inductive inquiry, which are expressed in five canons being:

Method of Agreement

A common example of this method occurs in determining the cause of an incidence of food poisoning within a group of people, where using the method of agreement, one will determine which food stuffs, each of the persons so affected have eaten. It is the search for a common item or instance which may be suspected as the probable cause.

Method of Difference

In contrast the method of difference exploits our knowledge of cases where, the incident does not occur by looking for differences in the surrounding circumstances.

Method of Indirect Difference

In its strict form the method of difference looks at just two incidents, the method of indirect difference combines this with the method of agreement to produce a much more powerful technique.

The quest for hidden information in complex databases as variously termed: *knowledge discovery, scientific discovery, scientific knowledge discovery, data mining, information mining, pattern discovery* or *data analysis,* is generally suggestive of an underlying notion that there is, in fact information and or knowledge to be found within databases, and that all one has to do is look for it. It should be borne in mind that there may well be nothing of interest or of use, 'hidden' within the data. Further, such methods may disclose such things as we may already know or have discounted as meaningless by other means.

A particular problem is that a given search method, finds only certain types of pattern. If a system is to satisfy the requirement of highlighting all interesting data, it must meet some criterion of completeness and find all of the patterns which are contained therein. Bacon and Mill both assert that their exposition of the process of discovery identifies all the methods needed to reveal the laws of nature. Even if this were not so coverage of Mill's Methods can provide a necessary, if not sufficient, condition for completeness. It is asserted here that the breadth of the methods is not achievable with one algorithm, and that any such system must be expandable or extendible to include a variety of search methods in order to find as many possibly patterns and types of patterns as may exist in the data.

So it is to a discussion of the different systems which have been proposed, to which we now turn, in an effort to gain an insight to the advantages and possible failings or weaknesses which they have.

3. Fundamentals

Bacon's exposition was at least in part aimed at opening a path between practical experience and unsupported theorising. Bacon asserted that physical inquiries advanced in one of two ways, both of which he condemned:

1. Introduction

The quest for knowledge about the world and what laws govern it, is not new. It is much older than computers and databases. The ancients speculated about such matters and influenced our conceptions of the world and in the enlightenment Sir Francis Bacon set out his Novum Organon.[1] Bacon stated eleven directions for "true interpretations of nature" which were to show "how to educe ..... axioms from experience" and in turn "how to deduce and derive new experiments from axioms". It should also be noted that Bacon asserted that the out come of such experiments would be hypotheses which in turn would need to be tested in the light of particulars.

At the time Bacon was criticised for attempting to reduce discovery to a logical or mechanistic process. Yet, at least on the face of it, serious consideration of the proposals made by Bacon and later formulated by Mill, in his System of Logic, have been largely ignored in the design of "intelligent" agents for data mining. It is this approach of inductive logic which we address, considering whether other workers in the field have followed the methods, whether it be by design or accident, and then give consideration to what aspects of Mill's formulation of Bacon's methods of inquiry have not been considered, which may be of use in the interpretation of data, for knowledge discovery.

2. Knowledge discovery and exploration

The value of data has only recently been noted in regard to lost information, which can cost an organisation dearly. This may be in the form of lost opportunities, such as inappropriately directed marketing or in persons not being targeted. What is not realised at large, is that many organisations in holding complex databases, hold within their systems, highly specific information, which could increase their effectiveness, possibly by orders of magnitude.

---

[1] Novum organum sive indicia de interpretatione nature. The New Organon or True Directions concerning the Interpretations of Nature.

Inductive Methods of Data Mining

Tremaine A. O. Cornish
Department of Computer Science, Brunel University, Uxbridge, Middlesex,
UB8 3PH, UK.
E-mail - Tremaine.Alvin.Cornish@brunel.ac.uk

Abstract - **The growth of databases, in both their size and complexity along with their ready access has precipitated the realisation that within such vast collections of data, there is probably a body of information and knowledge there to be discovered or discerned. It is proposed here that the essence of what has variously been termed: knowledge discovery, scientific discovery, and data mining, is an inductive process which was expounded by Sir Francis Bacon in 1620 and formulated by John Stuart Mill in 1843. Further, it is suggested that there are some approaches which have been proposed which go some way toward forfilling the Baconinan method of scientific discovery, though they were unaware of the similarities which are in evidence. Those aspects of Mill's Methods which are not addressed elsewhere are considered as a means toward, at least partial, resolution of the identified problem.**