Fuzzy Taxonomies for Creative Knowledge Discovery

Trevor Martin[1,2], Zheng Siyao[1,3] and Andrei Majidian[2]

[1] AI Group, University of Bristol, BS8 1TR UK
[2] Intelligent Systems Lab, BT, Adastral Park, Ipswich IP5 3RE, UK
[3] School of Computer Science and Engineering, BeiHang University, Beijing, China

Trevor. Martin@bristol.ac.uk, zhengsyao@gmail.com, Andrei .Majidian@bt.com

Abstract. A systematic form of creative knowledge discovery is outlined, requiring taxonomies to generalise knowledge structures and mappings between taxonomies to find parallels between knowledge structures from different domains. These share many of the features needed to handle uncertainty in the semantic web, and results will be relevant to the URSW community.

Keywords: fuzzy taxonomy, creative knowledge discovery, fuzzy association rules, uncertainty in semantic web

1 Introduction

Almost by definition, creative knowledge discovery is difficult to automate and harder to assess objectively. By creative knowledge discovery, we mean finding previously unknown links between concepts or small "chunks" of knowledge in such a way that useful additional knowledge is generated. It can be distinguished from "standard" knowledge discovery by defining the latter as the search for explanatory and/or predictive patterns and rules in large volume data within a specific domain. For example, a knowledge discovery process might examine an ISP(internet service provider)'s customer database and determine that people who have a high monthly spend and who send more than three emails to the support centre in a single month are very likely to change to a different provider in the following month. Such knowledge is implicit within the data but is useful in predicting and understanding behaviour.

By contrast, creative knowledge discovery is more concerned with "thinking the unthought-of" and looking for new links, new perspectives, etc. Such links are often found by drawing parallels between different domains and looking to see how well those parallels hold - for example, compare the ISP example mentioned above to a hotel chain finding that regular guests who report dissatisfaction with two or more stays often cease to be regular guests unless they are tempted back by special treatment (such as complimentary room upgrades). This is a simple illustration of similar problems (losing customers) in different domains. A solution in one domain (complimentary upgrades) could inspire a solution in the second (e.g. a higher download allowance at the same price). Of course, such analogies may break down when probed too far but they often provide the creative insight necessary to spark a

new solution through a new way of looking at a problem. In many cases, this inspiration is often referred to as "serendipity", or accidental discovery.

It is possible that many serendipitous discoveries are subsequently rationalised as the outcome of rigorous application of the scientific process. The traditional view of the scientist is as a generator and tester of hypotheses - often this is presented as an almost mechanical process and systems such as King's robot scientist [1] take this to an extreme, using an inductive logic programming approach to systematically generate and test hypotheses in a laboratory.

In this paper we outline a project to automate creative knowledge discovery. The aim is to find parallels between different knowledge repositories - in this case, semantically annotated networks of documents or process models - in the hope of transferring useful links from one network to another. In the case of process models from different domains, the aim is to identify possible improvements in one process if its analogue in the other domain is more efficient in some way.

This work shares many of the problems faced by research into uncertainty in the semantic web - the mapping between repositories is very similar to a mapping between ontologies, and the creation of knowledge networks encounters several issues that are well-known from the semantic web, such as the need for imprecise concepts, integration of sources that represent entities and classes at different levels of detail etc. The work is at an early stage, and this paper briefly outlines (i) a possible approach to automating creativity which relies on the use of fuzzy taxonomies and (ii) preliminary work on automatic extraction of taxonomies from data; this requires a representation of uncertainty similar to that needed for the semantic web.

2 A Method for Creative Knowledge Discovery

Can creativity - in this sense of suddenly making novel connections - be automated? Koestler [2] summarised this view of creativity as follows:
"The creative act is not an act of creation in the sense of the Old Testament. It does not create something out of nothing: it uncovers, selects, re-shuffles, combines, synthesizes already existing facts, idea, faculties, skills. The more familiar the parts, the more striking the new whole"

Table 1 - attributes of two music players (taken from [4])

Conventional tape recorder	Sony Walkman
big	small
clumsy	neat
records	does not record
plays back	plays back
uses magnetic tape	uses magnetic tape
tape is on reels	tape is in cassette
speakers in cabinet	speakers in headphones
mains electricity	battery

Sherwood [3] proposes a systematic approach, in which a situation or artefact is represented as an object with multiple attributes, and the consequences of changing

attributes, removing constraints, etc are progressively explored. For example, given an old style reel-to-reel tape recorder as starting point, Sherwood's approach is to list some of its essential attributes, substitute plausible alternatives for a number of these attributes, and evaluate the resulting conceptual design or solution. Table 1 shows how this could have led to the Sony Walkman in the late 70s [4]. Again, with the benefit of hindsight the reader should be able to see that by changing magnetic tape to a hard disk and considering the way music is purchased and distributed, the same method could (retrospectively, at least) lead one to invent the iPod. Of course, having the vision to choose new attributes and the knowledge and foresight to evaluate the result is the hard part - and the creative steps are usually only obvious with hindsight.

This systematic approach is ideally suited to handling data which is held in an object-attribute-value format, provided we have a means of changing/generalising attribute values. We intend to use taxonomies for this purpose, so that "sensible" changes can be made (e.g. *mains*, *battery* are both possible values for a *power* attribute). Representing an object O as a set of attribute-value pairs $\{(a_i, v_i) | \textit{attribute } a_i \textit{ of object O has value } v_i\}$ we generate a new "design" $O^* = \{(a_i, T(v_i))\}$ by changing one or more values using T_i, a non-deterministic transformation of a value to another value from the same taxonomy. Given sufficient time, this would simply enumerate all possible combinations of attribute values. We can reduce the search space by looking at the solution to an analogous problem in a different domain.

Our aim is to adapt previously developed tools for taxonomy matching [5] so that analogies can be found; the next section briefly outlines a way to extract taxonomic structure when it is not explicitly available.

3 Extracting Embedded Soft Taxonomies

An ontology essentially consists of a taxonomy of concepts, one or more relations between concepts, and rules which impose constraints and allow data transformation. The idea of an ontology is central to the semantic web [6], although there can be a very high cost in creation and maintenance. This is reflected in practical experience - it is rare to find web-based data that is fully marked up with RDF or OWL metadata. It is far more common to encounter data that is stored in a relational database or an equivalent XML-tagged format. Such data often contains implicit taxonomies - a relational table may flatten hierarchical data into one or more attributes. For example, a film database may record genre(s) and sub-genre(s) as separate fields, hiding the hierarchical dependency. The hierarchy may be obvious to a human reader of the data, but it is invisible to the machine. Similarly, XML tags can hide structure. XML relies on human interpretation for its "semantics" - a programmer can take advantage of the fact that <*iPod*> and <*walkman*> are subtypes of <*music Player*>, but a program has no way of knowing this unless it is made explicit by means of a taxonomy. Although a well-designed schema will make hierarchical structure explicit, our experience is that a significant proportion of data sources rely on programmer intuition instead.

We have investigated formal concept analysis (FCA) [7, 8] as a way of extracting hidden structure from a dataset in object-attribute-value form. In its simplest form, FCA considers a binary-valued table, where each row corresponds to an object and

each column to an attribute (property). The extension to a fuzzy case is (relatively) straightforward, by considering a fuzzy relation R^* and alpha-cuts which reduce the problem to the crisp case. A brief outline and promising initial results are given in [9].

4 Applications

Two specific domains form demonstrators for this work. XML process mining algorithms exist to discover process model from log files; various additions include heuristic and fuzzy approaches to handle noisy data. Semantic processing mining involves ontology knowledge. The ProM [www.processmining.org] platform takes SA-MXML (semantic annotated mxml) files as input, where the annotation conforms to the Web Service Modelling Language. The aim of this demonstrator is to find (partial) similarities between process models in different domains, and use process simulation tools to determine whether one process can be improved by slightly altering it to match the second process more closely. The second demonstrator is based on web forum discussions and support centre documentation, and will attempt to improve the automated provision of "help" information.

5 Summary

This paper has briefly outlined a project to automate aspects of creative knowledge discovery. The project is in early stages. Although not a direct application of uncertain reasoning in the semantic web, it shares many of the same problems and useful cross-fertilisation of ideas should be possible.

Acknowledgement : this work was partly funded by the FP7 BISON (Bisociation Networks for Creative Information Discovery) project, number 211898.

6 References

1.King, R.D., et al., *Functional genomic hypothesis generation and experimentation by a robot scientist*. Nature, 2004(6971): p. 247-251.
2.Koestler, A., *The Act of Creation*. 1964: Macmillan. 751.
3.Sherwood,D. *SilverBullet Machine::Guide to Creativity*,2009, www.silverbulletmachine.com
4.Sherwood, D., *Koestler's Law: The Act of Discovering Creativity-And How to Apply It in Your Law Practice*. Law Practice, 2006. **32**(8).
5.Martin,T.P, B.Azvine, Y.Shen, *Granular Assoc Rules for Multiple Taxonomies: A Mass Assignment Approach* in *Uncertain Reasoning in the Sem Web*, M.Nickles, Ed 2008,Springer.
6.Berners-Lee, T, J.Hendler, and O.Lassila, *Semantic Web*, in *Scientific American* 2001 p28-37.
7.Ganter, B, R. Wille, *Formal Concept Analysis: Mathematical Foundations*. 1998: Springer.
8.Priss, U., *Formal Concept Analysis in Information Science*.
9.Majidian, A. and T.P. Martin. *Extracting Taxonomies from Data - a Case Study using Fuzzy FCA*. in *Web Intelligence-09*. 2009. Milan, Italy: IEEE Computing.