TÍTULO: Princípios Teóricos das Árvores de Conhecimentos
AUTOR: Michel Authier

Michel Authier, criador das "Árvores de Conhecimentos", é Diretor de Pesquisas e Métodos, e *também Presidente do Conseil de Surveillance* da *Société Trivium S.A., em Paris.*

Apresentação

O conceito das "Árvores de Conhecimentos" foi elaborado em novembro e dezembro de 1991. Os princípios subjacentes a esta elaboração são ao mesmo tempo matemáticos, filosóficos e sociológicos. Não há qualquer razão para pensar que uma destas três dimensões tenha sido mais importante do que as demais. A evolução de cada uma permitiu ou suscitou avanços nas outras. Face ao desenvolvimento cada vez mais sofisticado das técnicas de análise de dados, nossa iniciativa prioriza a necessidade de uma técnica de síntese de dados para fazer emergir o conhecimento, o sentido, de um conjunto de informações complexas. Esta iniciativa, que visa instrumentalizar o domínio da complexidade, necessita a implicação do usuário (teoria moderna do observador); ao contrário de uma iniciativa analítica, que procura desvelar os arcanos de uma complicação extrema, que tenta privilegiar a independência do resultado em relação a quem o manipula. (teoria clássica da objetividade).

Neste texto nos interessaremos apenas pelos princípios matemáticos que permitem, a partir de um conjunto abundante de informações, estabelecer uma síntese e representá-la sob a forma de cartogramas. Se fosse preciso vincular a qualquer preço estes princípios a um quadro teórico preexistente, poderíamos dizer com alguma aproximação que estão situados no enquadre da teoria da agregação das preferências individuais ou teoria da utilidade coletiva.

O Problema

As soluções matemáticas subjacentes aos algoritmos propostos pela Trivium em Gingo estão ligadas ao seguinte problema muito geral:

- Seja um conjunto de registros (variáveis) e um conjunto de indivíduos (subconjuntos ordenados de brevês). Nosso conceito de indivíduo aproxima-se do construído por G. Simondon.

Existe uma relação de ordem sobre os brevês que não contrarie a ordem das listas?

Se não, como minimizar o enfraquecimento das informações contidas nas listas para obter uma ordem sobre o conjunto dos brevês?

A abordagem desenvolvida pela Trivium é uma resposta a este problema. O "espaço solução" da síntese das informações ligadas aos indivíduos pode ser chamado "cinemapa", o que a Trivium propõe chama-se "Árvores de Conhecimentos".

No final do século XVIII, Condorcet, em seus estudos das matemáticas sociais (por exemplo, em "sur la manière de connaître le voeu de la pluralité dans les élections" ou "Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix", encontra-se diante do problema da construção da opinião geral a partir das opiniões individuais. Compreende-se facilmente que o conjunto dos indivíduos corresponde a uma coleção de opiniões particulares e que a Árvore de Conhecimentos nada seria além da opinião geral pesquisada. Se o

problema é solucionado para duas pessoas, além disso, estamos diante de situações paradoxais. Em 1951, Arrow, em "Escolha social e valores individuais", verifica a impossibilidade de resolver, sem outras informações, o problema de Condorcet.

Estatística

Condorcet já havia percebido que para "construir" a opinião geral a partir das opiniões particulares, era preciso admitir um princípio exterior que resolveria as situações de conflitos entre as opiniões (é como o que funda o princípio majoritário). Por outro lado, ele faz uma crítica muito fina demonstrando que, de fato, seria preciso multiplicar os procedimentos de escolha para resolver seu problema. Outras abordagens serão propostas em seguida, apoiando-se em técnicas estatísticas cada vez mais sofisticadas (que podem servir para muitas outras coisas) como a análise fatorial de correspondência, a análise em componente principal, a análise de correspondências múltiplas, a análise fatorial múltipla, a análise fatorial relacional etc. Todos estes métodos necessitam de um bom domínio de técnicas matemáticas, e um importante trabalho de preparação dos dados deve ser realizado: ponderação, escolha de agrupamentos, tipologia de variáveis, de indivíduos, de modalidades... Também é necessário, para interpretar os resultados, dominar bem a inércia dos eixos e o sentido dos fatores, as fórmulas de transição entre as formas, as situações significativas (hierarquias, efeito Gutman etc...) Em todos os casos, os resultados não podem ser obtidos sem aguda perícia e sem uma numerização importante dos dados que permitiria, entre outras coisas, definir uma métrica sobre o conjunto inicial de informações. Neste tipo de abordagem (numerização *a priori*) a proximidade nada mais é do que a aplicação desta métrica.

Teoria dos jogos

Poder-se-ia também procurar na teoria dos jogos a origem das técnicas matemáticas das Árvores de Conhecimentos. Mesmo se uma Árvore de Conhecimentos pode, de algum modo, parecer como uma árvore de jogo, as técnicas utilizadas em teoria dos jogos (álgebra linear, análise matemática, programação linear, otimização, probabilidade etc.) são muito diferentes daquelas das Árvores de Conhecimentos. Por outro lado, uma Árvore de Conhecimentos nada tem a ver com uma "árvore de jogo", na medida em que mesmo se identificarmos a jogadores cada um dos indivíduos, nunca estaríamos numa situação análoga, pois não estamos numa situação em que os jogadores jogam cada um por sua vez mas, bem ao contrário, todos juntos (como os músicos de uma orquestra...). Sublinhemos de passagem que as Arvores de Conhecimentos nada têm a ver com as aproximações probabilísticas na medida em que, por definição, estas abordagens tendem a apagar os eventos raros em relação aos mais prováveis, enquanto que o que interessa particularmente às Árvores de Conhecimentos é fazer aparecer os sinais fracos que estão mais freqüentemente ligados a eventos raros, portadores de sentido.

Uma abordagem original: nem estatística, nem probabilista

As abordagens preconizadas pela Trivium dão as costas aos métodos que resumimos. A exemplo das abordagens científicas contemporâneas que renunciam à explicação dos fenômenos pelo conhecimento das trajetórias precisas dos elementos que deles participam, não procuramos obter uma solução pela síntese de causas associadas ao acontecimento de cada informação. Trata-se, para as Árvores de Conhecimentos, de exprimir a opinião geral de um coletivo de agentes, que (virtuais ou reais) exprimem as listas ordenadas. Para isso

escolhemos elaborar esta síntese não levando em conta qualquer indivíduo em particular, mas examinando-os todos juntos, e interessando-nos pelas etapas sucessivas desta expressão coletiva. Trata-se de uma aproximação sistêmica em que cada acontecimento está implicado no conjunto do sistema. Cada elemento é parte inseparável do complexo e não um pólo bem definido ligado inextricavelmente a uma rede de outros pólos. É por esta razão que toda variação sobre as informações iniciais é susceptível de agir sobre o conjunto da solução. Assim, à menor mudança no sistema de informação, o cálculo da síntese será integralmente refeito. Compreender-se-á então a importância de não fazer este cálculo depender de uma parametragem qualquer e de se estar seguro de sua prontidão.

Esta lógica, na qual a inferência proposicional é dissolvida, da qual nós não conhecemos equivalente em matemática, poderíamos chamá-la como `lógica quântica", assimilando cada indivíduo a uma "possibilidade" de indução. Em outras palavras, não procuramos uma solução que explicitaria com certeza quais são os lugares de dependência entre os registros, graças a uma análise dos indivíduos. Ao contrário, estabelecemos com precisão uma solução estável (quer dizer, invariante com o lugar e o tempo) que define sem contestação possível o estado das ligações entre os subconjuntos de registros (camadas) que devem ser os menores possíveis. A "lógica quântica" à qual fazemos alusão, por analogia com a física de mesmo nome, não procura efeitos de elementos particulares (as partículas), mas os de conjuntos (os pacotes de partículas) cujo comportamento se pode conhecer com precisão ainda que se admita o princípio da incerteza sobre os elementos.

Já que permanecemos sobre conjuntos finitos, a linguagem de expressão pode ser indiferentemente a dos hipergrafos, a da topologia, a das estruturas de ordem. O espaço-solução, por sua própria existência (e não o inverso), estabelece uma proximidade que permite responder rapidamente a interrogações do tipo "quais são as *n* mais próximas de...?"; ele é então um espaço topológico sem métrica *a priori*, mesmo se *a posteriori* é possível estabelecer uma, já que a topologia estará bem evidentemente separada. Sob este aspecto, e por causa do modo recorrente de elaboração da solução, poderíamos também chamar esta técnica de "topologia recursiva", já que a topologia do espaço-solução se constrói recursivamente. É, aliás, esta recursividade que garante a possibilidade de uma solução informática

Princípio de otimização

De fato, o que procuramos através da elaboração deste espaço topológico é a exibição de uma forma que dê um sentido o mais de acordo possível com o conjunto das informações apresentadas no conjunto dos indivíduos. Isto claramente significa que a projeção de cada indivíduo na Árvore de Conhecimentos deve permitir reencontrar pelo menos as informações sustentadas pelos indivíduos. É este princípio que o algoritmo deve respeitar para gerar o "espaço-solução".

Recusando qualquer métrica ou princípio de ponderação como condição de elaboração dasolução, a teoria dos "cinemapas" permite reagir muito depressa a toda modificação do conjunto das informações iniciais. De uma certa maneira, a finalidade da representação é respeitar o seguinte princípio: as opiniões "próximas" (quer dizer, pouco divergentes) devem ter representações "próximas" (no sentido induzido pela topologia) no espaço-solução (a Árvore de Conhecimentos). Contudo, esta noção de proximidade é mais potente do que é habitualmente admitido. De fato ela permite que se faça uma idéia das relações

entre duas expressões integrando as relações entre todas as outras expressões (referencial contextualizado ou 'relativista'), e não sobre um único exame da distância entre estas duas expressões (referencial absoluto).

É bem evidente que para respeitar o princípio de otimização, as camadas devem ser o mais numerosas possível e, entretanto, serem mínimas (por exemplo, se não houvesse qualquer contradição entre os indivíduos, as camadas se reduziriam cada uma a um só elemento, e a Árvore de Conhecimentos seria topologicamente equivalente a uma coluna : conjunto de pontos totalmente ordenados)

Teoremas originais provam que as camadas formam uma partição do conjunto dos registros, que as ligações entre os componentes conexos contidos nas camadas formam uma arborescência, que os indivíduos induzem uma estrutura fina sobre as camadas, que o conjunto destas informações pode ser representado numa imagem 2D, topologicamente equivalente a uma árvore.

Comentários

A estrutura de árvore não é então uma finalidade que condiciona o tratamento da informação pelos algoritmos, ela é a resultante do tratamento pelo algoritmo que procura produzir uma forma que contrarie o menos possível as formas particulares e triviais (as colunas) induzidas pelos indivíduos. Em nenhum caso a forma "árvore" determina o funcionamento do algoritmo. Além disso, esta forma não induz que exista uma estrutura arborescente entre os registros já que em certos componentes conexos as camadas podem representar diversos registros que entre si não são estruturados de maneira arborescente.

Outros teoremas permitem reduzir a complexidade dos cálculos à linearidade em função do número de indivíduos, e em n*log (n), n sendo o número de registros. Um algoritmo de tipo totalmente diferente gera as posições de cada registro a fim de gerar a representação. A mesma fórmula permite evidentemente re-situar em "tempo real" cada registro, à menor variação do sistema de informações.

Evidentemente, nenhum elemento estatístico participa da estruturação da representação. Ao contrário, é possível exprimir o quantitativo, em particular através da "coloração" que pode tomar cada elemento da representação.

Para concluir, é preciso sublinhar que a existência de Árvores de Conhecimentos ligadas a sistemas de informações diferentes conduz ao estudo sistemático de uma estrutura sobre o conjunto das "Árvores de Conhecimentos", em que podem ser definidas operações : soma, diferença, dualidade... Por exemplo, o conceito de dualidade, tão rico, onipresente em análise de dados, encontra aqui toda a sua riqueza e sua flexibilidade, na medida em que as ligações entre os registros e os indivíduos podem perfeitamente se inverter. Assim, por exemplo, é perfeitamente indiferente fazer a Árvore de Conhecimentos dos produtos sobre os quais um certo número de indivíduos exprimiu sua preferência ou fazer a dos indivíduos que exprimiram sua preferência a respeito de um certo número de produtos.

Se a informação estatística não participa da elaboração da árvore, isto não significa que ela esteja ausente da representação. Todas as informações estatísticas são diretamente acessíveis e, além disso, elas se tornam visíveis pelos efeitos de coloração dos elementos da representação. A analogia com a cartografia aqui é perfeita. A forma de um mapa nada diz do relevo do território (que a coloração pode representar), ela nada informa a não ser em relação a um certo nível (geralmente o do mar). Para as árvores, o análogo deste nível será o grau de restrições aplicadas à estruturação.

Como é facilmente visível, este algoritmo submete as informações a duas fases de estruturação:

A identificação de camadas graças ao princípio de "geração".

A identificação de componentes conexos de camadas, graças ao princípio de "conexidade".

Estes dois princípios aplicados em toda sua "pureza" matemática podem muito bem fazer aparecer apenas estruturas triviais ou totalmente disparatadas no caso em que as informações forem, ou fortemente repetitivas ou totalmente contraditórias. Neste caso, é de fato possível diminuir as restrições impostas por um ou outro dos princípios introduzindo os parâmetros no algoritmo (função moduladora). No caso, é possível obter uma forma mais legível e portanto fazer emergir um sentido mais explícito. Mas isto é obtido ao preço de uma margem de incerteza mais significativa sobre a maneira pela qual o algoritmo respeita a informação que ele trata. Esta margem de incerteza é numerizada; pode-se então afirmar que os instrumentos de regulagem do algoritmo permitem mensurar o delta de variação que existe entre uma forma que dá sentido a um conjunto de informações e este mesmo conjunto de informações

A noção de incerteza é fundamental para compreender a iniciativa científica e a prática das "Árvores de Conhecimentos". De fato, a condição para obter uma solução não trivial é que alguns registros sejam partilhados por certos indivíduos. Uma identificação totalmente precisa de um registro associado a um indivíduo nos conduziria a concebê-lo como de fato discernível (princípio dos indiscerníveis de Leibniz). A crença no fato de que um conhecimento pode ser extraído a partir de um conjunto de informações implica a possibilidade de elaborar sentido a partir deste sistema de informação, portanto de admitir a partilha de certos registros e, então, de aceitar um princípio de incerteza sobre as determinações das características dos registros. Para dar um exemplo, não se pode elaborar o conhecimento sobre as competências a não ser admitindo que a identificação exata da competência de uma pessoa não tem estritamente qualquer sentido!

Algumas vantagens da nossa abordagem

A preparação dos dados pode ser limitada ao mínimo, na medida em que não é absolutamente necessário ponderar as variáveis que caracterizam os indivíduos.

A dualidade (indivíduo/variável) é muito fácil de ativar, já que a ausência de ponderação dispensa cálculos de transposição.

A contextualização não coloca qualquer problema já que, por definição, ela é o próprio espaço de representação da solução: quer dizer, a Árvore de Conhecimentos. De fato, não há espaço (afim) preliminar à representação da síntese; não há então qualquer referencial absoluto no qual viriam se posicionar os eixos particulares. O espaço topológico das variáveis é o produto das interações entre todos os indivíduos; ele não preexiste a qualquer um deles e varia com cada variação. Isso torna totalmente vazia de sentido a noção de distância entre dois indivíduos independentemente dos outros. Mais uma vez, a contextualização se funda sobre o abandono da idéia preconcebida de um referencial absoluto. Espaço e indivíduo se definem, se influenciam mutuamente, estruturalmente isto não está longe de assemelhar-se à relatividade.

A ausência total de métrica preliminar permite aos sinais fracos (estatísticamente raros, mas estruturalmente coerentes) não desaparecer face a sinais

massivamente repetidos. Assim fenômenos mais raros podem aparecer muito visivelmente na representação se eles não forem redutíveis a fenômenos mais massivos.

A existência de um espaço estabelecido topologicamente e não metricamente não impede, em absoluto, que se faça emergir uma métrica deste espaço; ainda são possíveis cálculos sobre as proximidades. Além disso, o espaço poderá perfeitamente suportar (tornar visíveis) todas as informações estatísticas sobre as variáveis. Em particular, tudo o que diz respeito às lógicas de uso; daí a possibilidade de integrar em tempo real os efeitos de exploração do sistema de informações pelos usuários. Assim, em alguns minutos, se pode simular, por dezenas de conseqüências, hipóteses de transformação do sistema de informações.

A rapidez de reação do instrumento às variações do sistema é o que permite o domínio da complexidade; sem ela não haveria meio de reagir num tempo suficiente à evolução constante da realidade que nos rodeia.

Está bem claro que este nível de desempenho pode ser atingido graças ao progresso dos sistemas informáticos mas, mais ainda, graças à extrema simplicidade da estrutura dos dados de início e à fraca complexidade da algorítmica implementada.