Localizando Padrões

Padrões são unidades de informação que se repetem. A tarefa de localizar padrões não é privilégio da Mineração de dados. Nosso cérebro utiliza-se de processos similares, pois muito do conhecimento que temos em nossa mente é, de certa forma, um processo que depende da localização de padrões. Para exemplificar esses conceitos, vamos propor um breve exercício de uma indução de regras abstratas. Nosso objetivo é tentar obter alguma expressão genérica para a seguinte seqüência:

Seqüência original: **ABCXYABCZKABDKCABCTUABEWLABCWO**
Observe atentamente essa seqüência de letras e tente encontrar alguma coisa relevante. Veja algumas possibilidades:

Passo 1: A primeira etapa é perceber que existe uma seqüência de letras que se repete bastante. Encontramos as seqüências "AB" e "ABC" e observamos que elas ocorrem com freqüência superior à das outras seqüências.

Passo 2: Após determinarmos as seqüências "ABC" e "AB", verificamos que elas segmentam o padrão original em diversas unidades independentes:

"ABCXY"
"ABCZK"
"ABDKC"
"ABCTU"
"ABEWL"
"ABCWO"
Passo 3: Fazem-se agora induções, que geram algumas representações genéricas dessas unidades:

"ABC??" "ABD??" "ABE??" e "AB???",
onde '?' representa qualquer letra

No final desse processo, toda a seqüência original foi substituída por regras genéricas indutivas que simplificou (reduziu) a informação original a algumas expressões simples. Esta explicação é um dos pontos essenciais da mineração de dados, como se pode fazer para extrair certos padrões de dados brutos. Contudo, mais importante do que simplesmente obter essa redução de informação, esse processo nos permite gerar formas de predizer futuras ocorrências de padrões.

Exemplo Prático

Vamos observar aqui apenas um pequeno exemplo prático do que podemos utilizar com as expressões abstratas genéricas que obtivemos. Uma dessas expressões nos diz que toda vez que encontramos a seqüência "AB", podemos inferir que iremos encontrar mais três caracteres e isto completaria um "padrão". Nesta forma abstrata ainda pode ficar difícil de perceber a relevância deste resultado. Por isso vamos usar uma representação mais próxima da realidade.

Imagine que a letra 'A' esteja representando um item qualquer de um registro comercial. Por exemplo, a letra 'A' poderia significar "aquisição de pão" em uma transação de supermercado. A letra 'B' poderia, por exemplo, significar "aquisição de leite". A letra 'C' é um indicador de que o leite que foi adquirido é do tipo desnatado. É interessante notar que a obtenção de uma regra com as letras "AB" quer dizer, na prática, que toda vez que alguém comprou pão, também comprou leite. Esses dois atributos estão associados e isto foi revelado pelo processo de descoberta de padrões.

Esta associação já nos fará pensar em colocar "leite" e "pão" mais próximos um do outro no supermercado, pois assim estaríamos facilitando a aquisição conjunta desses dois produtos. Mas a coisa pode ir além disso, bastando continuar nossa exploração da indução.

Suponha que a letra X queira dizer "manteiga sem sal", e a letra 'Z' signifique "manteiga com sal". A letra 'T' poderia significar "margarina". Parece que poderíamos tentar unificar todas essas

letras através de um único conceito, uma idéia que resuma uma característica essencial de todos esses itens. Introduzimos a letra 'V', que significaria "manteiga/margarina", ou "coisas que passamos no pão". Fizemos uma indução orientada a atributos, substituímos uma série de valores distintos (mas similares) por um nome só.

Ao fazer isso estamos perdendo um pouco das características dos dados originais. Após essa transformação, já não sabemos mais o que é manteiga e o que é margarina. Essa perda de informação é fundamental na indução e é um dos fatores que permite o aparecimento de padrões mais gerais.

Qual a vantagem de assim proceder? Basta codificar a seqüência original substituindo a letra V em todos os lugares devidos. Assim fica essa seqüência transformada:

ABCVYABCVKABDKCABCVUABEWLABCVO
Daqui, o sistema de Mineração de Dados irá extrair, entre outras coisas, a expressão "ABCV", que irá revelar algo muito interessante:

A maioria dos usuários que adquiriram pão e leite desnatado
também adquiriram manteiga ou margarina.
De posse desta regra, fica fácil imaginar uma disposição nas prateleiras do supermercado para incentivar ainda mais este hábito. Em linguagem mais lógica, pode-se dizer que pão e leite estão associados (implicam) na aquisição de manteiga:

Pão, Leite => Manteiga

Fonte: Wikipedia