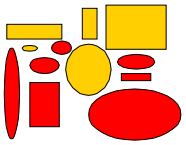
Agrupamento de Dados

Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação
Rodrigo Fernandes de Mello
mello@icmc.usp.br

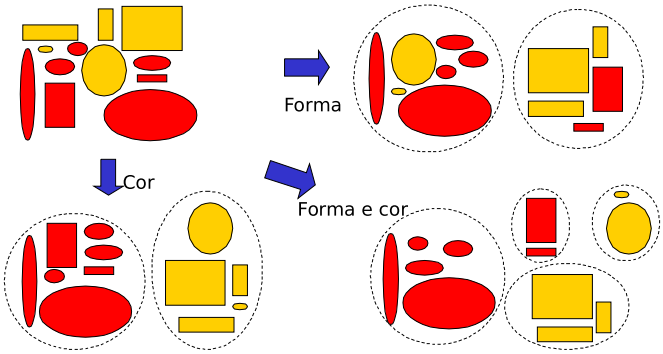
Agrupamento de Dados

- Segue o paradigma de aprendizado não-supervisionado
- Extração de conhecimento sem utilizar informações sobre as classes dos exemplos
- Busca organizar um conjunto de objetos em grupos
  - De acordo com medidas de similaridade ou dissimilaridade



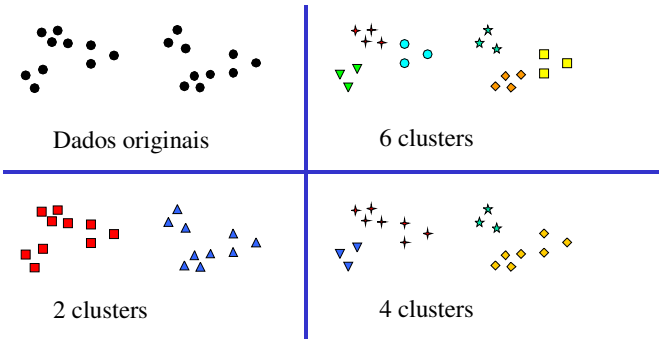
Como organizar?

Agrupamento de Dados



Questões Relevantes

- Quantos grupos (ou clusters)?



Questões Relevantes

- Quais formatos de clusters?

Agrupamento de Dados

- Definição:
  - Considere elementos identificados pelo conjunto:

    $S = \{1, 2, ..., n\}$
  - Considere uma função de dissimilaridade $d : S \times S \rightarrow R$ tal que para todo i, j pertencente a S:

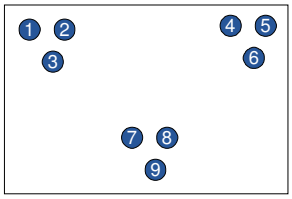
    $d(i, j) \geq 0$

    $d(i, j) = 0$, se $i = j$

    $d(i, j) = d(j, i)$
  - Uma função de agrupamento f é uma função que recebe d e retorna uma partição Γ de S
  - Um elemento pode:
    - Pertencer a somente um grupo (crisp)
    - A vários grupos com diferentes graus de pertinência (fuzzy)

Agrupamento de Dados: Crisp

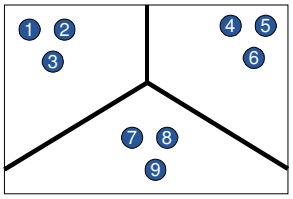
- Um exemplo visual:



- S = {1, 2, 3, 4, 5, 6, 7, 8, 9}

Agrupamento de Dados: Crisp

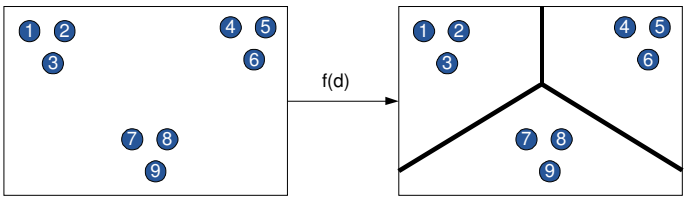
- Um exemplo visual:



- Γ = {{1, 2, 3}, {4, 5, 6}, {7, 8, 9}}
- Neste caso há três grupos:
  - Grupo 1 = {1, 2, 3}
  - Grupo 2 = {4, 5, 6}
  - Grupo 3 = {7, 8, 9}

Agrupamento de Dados: Crisp

- Logo o objetivo de Algoritmos de Agrupamento de Dados é:



- As funções mais comuns de distância adotadas consideram:
  - Vizinhança ou Proximidade
    - Ex: K-means e RBF
  - Densidade
    - Ex: DBScan

Agrupamento de Dados: Fuzzy

- Γ = { {(1, 1.0), (2, 1.0), (3, 0.9)},
  {(3, 0.1), (4, 0.6), (5, 1.0), (6, 1.0)},
  {(4, 0.4), (7, 1.0), (8, 1.0), (9, 1.0)} }
- Neste caso há três grupos:
  - No entanto cada elemento apresenta sua pertinência relativa ao grupo
    - (identificador do elemento, pertinência ao grupo)

Abordagens para Agrupamento de Dados

- Busca exaustiva
  - Verificar todos os possíveis agrupamentos de tamanho k para vários valores de k
    - Em que k representa o número de grupos
  - Abordagem de custo proibitivo para grande número de elementos
- Particional
  - Protótipos
  - Densidade
- Hierárquicos
- Baseados em otimização de função de custo
- Baseados em grafos

Agrupamento de Dados: Algoritmos Particionais

- Produzem um único agrupamento
- A maioria utiliza uma abordagem gulosa (greedy):
  - Escolha da melhor alternativa atual, sem considerar consequências futuras
  - Uma vez tomada a decisão, ela não é mais alterada
  - Geralmente resultados dependem da ordem em que elementos são apresentados ao algoritmo

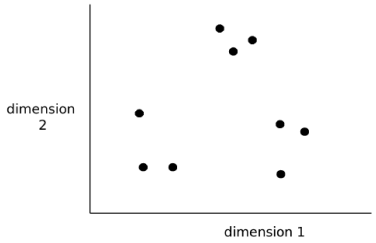
Agrupamento de Dados: Algoritmos Particionais

- Exemplos de algoritmos particionais:
  - K-médias
  - K-médias ótimo
  - K-médias seqüencial
  - SOM
  - FCM
  - DENCLUE
  - CLICK
  - CAST
  - SNN

Slide baseado nas notas de aula dos Profs. André C.P.L.F. de Carvalho e Ricardo J.G.B. Campello

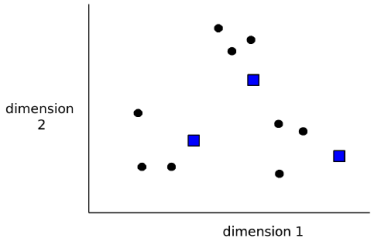
K-Means

- Provavelmente o algoritmo mais conhecido para agrupamento de dados
  - Busca particionar n objetos em k grupos, em que k < n
  - Objetos são associados ao grupo mais próximo
    - Função de distância considera proximidade
- Considere, por exemplo (com k = 3):



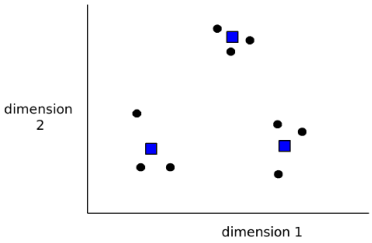
K-Means

- Inicialmente selecionamos k pontos aleatórios sobre o espaço
  - Esses pontos são também chamados de centróides ou protótipos
- Para cada objeto:
  - Computamos o centróide mais próximo e rotulamos esse objeto como associado a tal protótipo



K-Means

- Em seguida, recalculamos a posição dos centróides com base na posição de seus objetos associados



K-Means

- Para recalcular a posição de um centróide c consideramos a distância média de seus objetos relacionados:

$$avg_d = \frac{1}{p}\sum_{j=1}^{p} a_d$$

  - Em que:
    - a_d representa um atributo
    - p é o número de objetos associados ao centróide c
  - Em resumo:
    - Cada atributo do centróide recebe a média dos atributos dos objetos a ele associados
- A cada passo do algoritmo:
  - Centróides são movidos em direção a seus objetos associados
  - O algoritmo para quando não houver mais variações nos centróides
  - Ao final obtemos as coordenadas dos centróides que particionam o espaço

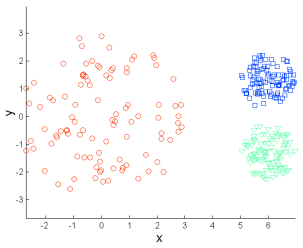
K-Means

- Limitações:
  - Escolha do valor para k
  - K-means tem problemas quando os grupos têm:
    - Formatos não hiperesféricos
    - Quando dados apresentam outliers
      - Esses influenciam os protótipos

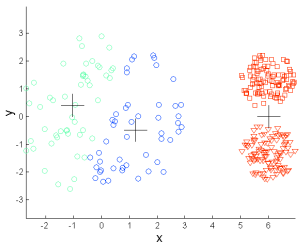
Slide baseado nas notas de aula dos Profs. André C.P.L.F. de Carvalho e Ricardo J.G.B. Campello

K-Means

- Exemplo:



Dados originais

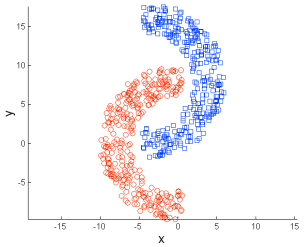


K-médias (3 Clusters)

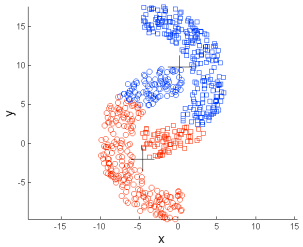
Slide baseado nas notas de aula dos Profs. André C.P.L.F. de Carvalho e Ricardo J.G.B. Campello

K-Means

- Exemplo:



Dados originais



K-médias (2 Clusters)

Slide baseado nas notas de aula dos Profs. André C.P.L.F. de Carvalho e Ricardo J.G.B. Campello

K-Means

- Implementação:
  - Implementar o Algoritmo K-Means e aplicá-lo sobre conjuntos de dados como:
    - Iris – http://archive.ics.uci.edu/ml/datasets/Iris
    - Wine – http://archive.ics.uci.edu/ml/datasets/Wine

Radial Basis Function

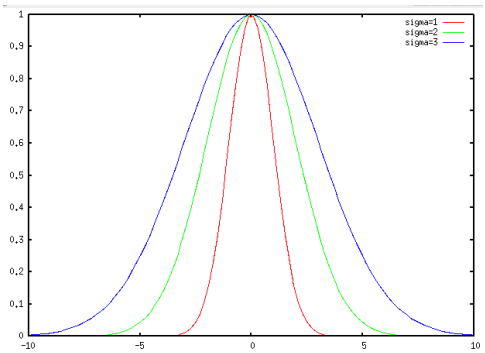
- Podemos utilizar Funções de Base Radial para agrupar dados
  - Essas funções consideram um raio de ativação ao redor de um centróide tal como:

$$act(i) = \exp^{-\frac{d(i,c)^2}{2.0 \cdot \sigma^2}}$$

  - Em que consideramos:
    - A distância d de um exemplo i ao centróide c
    - Sigma representa a abrangência da função radial

Radial Basis Function

- Conforme o valor de sigma, temos uma abrangência para a radial



Radial Basis Function

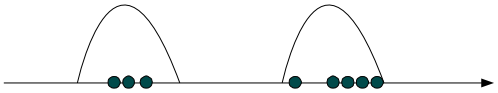
- O algoritmo funciona da seguinte maneira:
  - Em um primeiro momento nenhum dado foi observado e nenhuma função radial existe no sistema
  - Logo em seguida chega um exemplo para o algoritmo e ele cria uma função radial centrada nesse exemplo:

Radial Basis Function

- O algoritmo funciona da seguinte maneira:
  - Suponha a chegada de outro exemplo e criação de uma nova radial, pois nenhuma das radiais existentes (no caso somente uma) é capaz de representar tal exemplo

Radial Basis Function

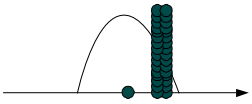
- O algoritmo funciona da seguinte maneira:
  - Quando um exemplo é recebido pelo algoritmo:
    - Verificamos se uma das radiais existentes é capaz de representá-lo
      - Para isso computamos a equação da radial (como a anteriormente vista) e verificamos se sua ativação é maior que um threshold ou limiar mínimo para ativação
      - Caso positivo, ocorre algo como abaixo...
    - Caso não seja, criamos uma nova radial centrada exatamente no exemplo recebido



- Ao final temos cada radial representando um grupo com seus objetos associados

Radial Basis Function: Centróides Adaptativos

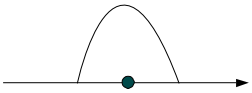
- O que ocorre se uma radial é formada sobre uma observação e depois vários outros exemplos são recebidos em regiões distantes desse centro?



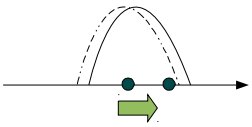
- Nesta situação:
  - Um centróide foi definido, no entanto vários outros exemplos “caíram” mais distantes do centro
  - Logo os dados estão mal distribuídos dentro desse grupo, o qual é representado por uma radial
  - Podemos mover o centro dessa radial para melhor representar os dados...

Radial Basis Function: Centróides Adaptativos

- Podemos mover o centro dessa radial para melhor representar os dados
  - Por exemplo, considere o cenário inicial:

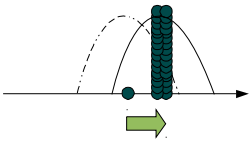


  - Ao receber outro exemplo, o centro da radial é alterado:



Radial Basis Function: Centróides Adaptativos

- Após receber os mesmos exemplos apresentados anteriormente, temos uma melhor representação para a radial



- Assim o centro da radial resultante é mais representativo, funcionando como um valor esperado de exemplos típicos

Radial Basis Function: Centróides Adaptativos

- Para realizar tal movimentação de centros podemos utilizar uma média móvel exponencialmente ponderada na forma:

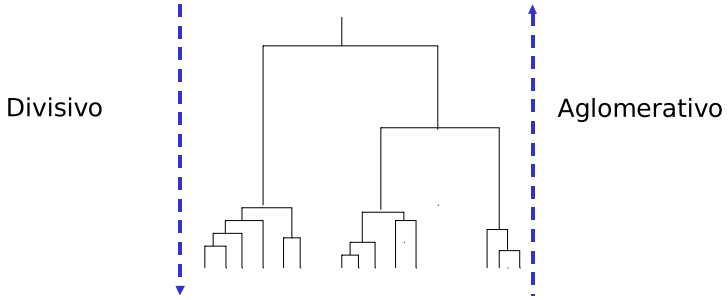
$$c_{t+1} = (1.0 - \alpha) \cdot c_t + \alpha \cdot i$$

- Em que consideramos:
  - Um exemplo i e um centro c
  - O centro c é adaptado no tempo (após receber um exemplo)
  - Uma taxa de adaptação alpha

Agrupamento de Dados: Algoritmos Hierárquicos

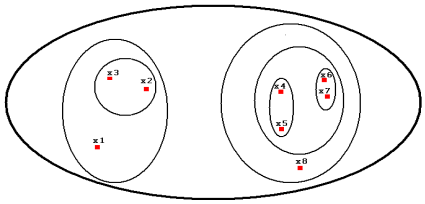
- Utilizam dendrogramas (diagrama em árvore)
  - Produz uma sequência (hierarquia) de agrupamentos
  - O corte de um dendrograma em qualquer nível produz uma simples partição
- Tipos:
  - **Aglomerativos:** combinam, repetidamente, dois clusters em um
    - A cada passo, combina os dois clusters mais similares
  - **Divisivos:** Dividem, repetidamente, um cluster em dois
    - A cada passo, divide o cluster menos homogêneo em dois novos clusters

Agrupamento de Dados: Algoritmos Hierárquicos



Agrupamento de Dados: Algoritmos Hierárquicos

- Ou Diagrama de Venn



Como computar a similaridade ou dissimilaridade entre elementos?

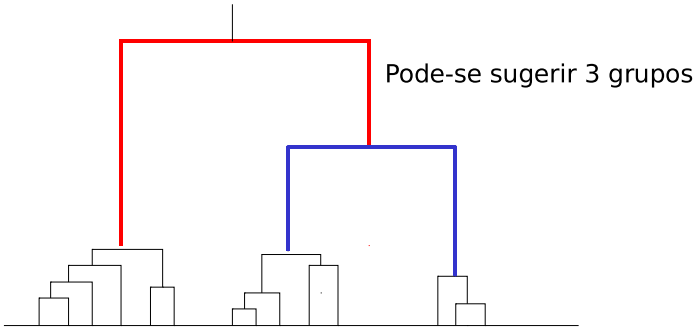
- Existem várias métricas:
  - Distância Euclidiana
  - Distância Manhattan (bloco-cidade)
  - Distância quadrática
  - Distância de Mahalanobis
  - Dynamic Time Warping
  - NCD, etc.

Como escolher uma partição?

- Como escolher partição?
  - Selecionando partição com *n* clusters na seqüência de agrupamentos da hierarquia
  - Partição que melhor se encaixa nos dados
    - Conhecemos os dados?
    - Temos uma métrica para avaliar?
  - Procurar no dendrograma grandes mudanças em níveis adjacentes
    - Nesse caso, uma mudança de *j* para *j-1* grupos pode indicar que *j* é o melhor número de grupos
    - Existem outros procedimentos, alguns mais objetivos

Como escolher uma partição?

- Exemplo:



Validação de agrupamentos

- Existem várias medidas para avaliar a qualidade de classificadores
  - Acurácia, precisão, revocação, F1
- Como avaliar os clusters gerados por um algoritmo de agrupamento?
- Por que avaliar agrupamentos?

Slide obtido das notas de aula dos Profs. André C.P.L.F. de Carvalho e Ricardo J.G.B. Campello

Validação de agrupamentos

- Por que avaliar agrupamentos?
  - Para evitar encontrar padrões em ruídos
  - Para comparar algoritmos de agrupamento
  - Para comparar duas partições
  - Para comparar grupos

Slide obtido das notas de aula dos Profs. André C.P.L.F. de Carvalho e Ricardo J.G.B. Campello

Medidas de validação

- Existem várias medidas de validação
  - Julgam aspectos diferentes
- Podem ser divididas em três grupos:
  - Índices ou critérios internos
    - Medem a qualidade da partição obtida sem considerar informações externas
  - Índices ou critérios externos
    - Medem o quanto os rótulos dos grupos casam com a classe verdadeira
  - Índices ou critérios relativos
    - Usados para comparar duas partições ou grupos

Slide obtido das notas de aula dos Profs. André C.P.L.F. de Carvalho e Ricardo J.G.B. Campello

Medidas internas

- Coesão de clusters
  - Mede o quão relacionados estão os objetos dentro de um cluster
- Separação de clusters
  - Mede quão distintos ou separados um cluster é dos demais clusters

Slide obtido das notas de aula dos Profs. André C.P.L.F. de Carvalho e Ricardo J.G.B. Campello

Exemplo

- Usando soma dos erros quadráticos (SSE)
  - Coesão é medida pelo SSE (Sum of Squared Error) dentro dos clusters

$$\mathbf{SSE} = \sum_{\mathbf{i}} \sum_{\mathbf{x} \in \mathbf{C_i}} (\mathbf{x} - \mu_\mathbf{i})^\mathbf{2}$$

  - Separação é medida pelo BSS (Between Sum of Squares) entre clusters

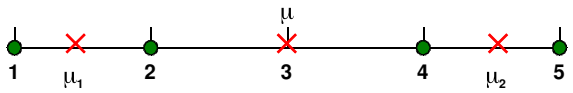
$$\mathbf{BSS} = \sum_{\mathbf{i}} |\mathbf{C_i}| (\mu - \mu_\mathbf{i})^\mathbf{2}$$

  - SSE + BSS = constante

$|C_i|$ é o número de elemento contidos no cluster C_i

Slide baseado nas notas de aula dos Profs. André C.P.L.F. de Carvalho e Ricardo J.G.B. Campello

Exemplo



K=1 cluster:

$$\text{SSE} = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$
$$BSS = 4 \times (3-3)^2 = 0$$
$$Total = 10 + 0 = 10$$

K=2 clusters:

$$\text{SSE} = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$
$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$
$$Total = 1 + 9 = 10$$

Slide obtido das notas de aula dos Profs. André C.P.L.F. de Carvalho e Ricardo J.G.B. Campello

Medidas internas

- Silhueta
  - Combina coesão com separação
  - Calculada para cada objeto que faz parte de um agrupamento
    - Baseada na proximidade entre os objetos de um cluster e na distância dos objetos de um cluster ao cluster mais próximo
  - Mostra quais objetos estão bem situados dentro dos seus clusters e quais estão fora do cluster apropriado

Medidas internas

- Silhueta
  - Para cada objeto *i*
    - *a(i)* = distância média de *i* aos demais objetos no mesmo cluster
    - *b(i)* = min (distância média de *i* aos objetos dos outros clusters)

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

  - Valores:
    - Entre -1 e +1
    - Quanto mais próximos de +1, melhor

Medidas externas

- Medidas orientadas a similaridade
  - Comparam duas partições
    - Índice Rand
    - Jackard

Medidas externas

- Índice Rand
  - Dado um conjunto S com n elementos e duas partições de S, tem-se:
    - f_{00} = número de pares de objetos com classes e clusters diferentes
    - f_{01} = número de pares de objetos com classes diferentes e mesmo cluster
    - f_{10} = número de pares de objetos com mesma classe e clusters diferentes
    - f_{11} = número de pares de objetos com mesmas classes e clusters

$$R = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

Medidas externas

- Índice Jackard
  - Dado um conjunto S com n elementos e duas partições de S, tem-se:
    - f_{00} = número de pares de objetos com classes e clusters diferentes
    - f_{01} = número de pares de objetos com classes diferentes e mesmo cluster
    - f_{10} = número de pares de objetos com mesma classe e clusters diferentes
    - f_{11} = número de pares de objetos com mesmas classes e clusters

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Medidas externas

- Adjusted Rand Index
  - Seja o conjunto S com N elementos e duas partições:

$$U = \{U_1, U_2, \ldots, U_R\} \text{ e } V = \{V_1, V_2, \ldots, V_C\}$$

  - Sobreposições de U e V são analisadas por uma tabela de contigência em que n_{ij} denota o número de objetos comuns entre grupos U_i e V_j, ou seja:

$$n_{ij} = |U_i \cap V_j|$$

U → partição gerada
V → partição verdadeira

U\V	V_1	V_2	$\cdots$	V_C	Sums
U_1	n_{11}	n_{12}	$\cdots$	n_{1C}	a_1
U_2	n_{21}	n_{22}	$\cdots$	n_{2C}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
U_R	n_{R1}	n_{R2}	$\cdots$	n_{RC}	a_R
Sums	b_1	b_2	$\cdots$	b_C	

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{N}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{N}{2}}$$

Medidas externas

- Adjusted Rand Index
  - Seu valor máximo é 1

Implementação

- Implementação:
  - Implementar a RBF tradicional
  - Implementar a RBF adaptativa
  - Agrupar dados provenientes de uma stream de áudio a fim de compactar essas dados
    - Verifiquemos os tamanhos de arquivo resultantes
- Exercício:
  - Buscar por um dataset de imagens de faces e realizar o agrupamento
  - Buscar por um dataset de músicas e realizar o agrupamento
    - Características extraída de áudio usando Mel Frequency Cepstrum Coefficients (MFCCs)

Referências

- Jon Kleinberg, An Impossibility Theorem for Clustering, 2002
- Albertini e Mello, A Self-Organizing Neural Network for Detecting Novelties, Proceedings of the 2007 ACM Symposium on Applied Computing, 2007