Research Directions in Sensor Data Streams: Solutions and Challenges

Eiman Elnahrawy - DCIS-TR-527,May 2003.

Eduardo Habib B. Maia

Daniel Câmara

{habib, danielc}@dcc.ufmg.br

2/11/2004

Referência

- Eiman Elnahrawy, *Research Directions in Sensor Data Streams: Solutions and Challenges*, DCIS Technical Report DCIS-TR-527, Rutgers University, May 2003.
- Cópia on-line : http://www.cs.rit.edu/~dmrg/adm_spring/reading/sensor/eiman03sensor_survey.pdf

Roteiro

- Introdução
- Background
- In-Network Storage
- In-network Aggregation
- Arquiteturas para Streams de Sensores
- Dados errados
- Conclusão

Introdução

- Considerando grandes redes com diferentes tipos de sensores
  - Telefones celulares, PDAs, veículos
- Data stream : produção contínua de dados

Introdução

- Trabalhos atuais
  - Focam em: restrições das redes de sensores
- Necessidades futuras
  - Mineração de dados on-line
  - Pesquisa e eliminação de ruídos em dados com ruídos
  - Determinação de *outliers*
  - Manipulação de dados incompletos

Introdução

- Motivação para estudos nestas áreas
  - Utilização real das redes de sensores
  - Manipular dados errados
    - 10% dos links sofrem com mais que 50% de taxa de perda no link
    - 5-10% de nodos falham a cada recuperação de dados

Background

- Limitações e problemas
  - Energia
  - Banda (1-100Kbps)
  - Armazenamento (8KB mem. Prog. 512B Mem. dados)
  - Processamento (4 MHz)
  - Processamento é ordens de grandeza menor que comunicação
  - Dados incompletos e imprecisos
    - Perda de pacotes e mudanças topológicas

Background

- Dados incompletos
  - Soluções de baixo nível (retransmissão)
  - Soluções customizadas de acordo com a aplicação
  - Faltam ainda soluções genéricas

Background

- Dados imprecisos
  - Tempo de *update* das bases de dados
    - Grande volume de dados
    - Banda e energia limitadas
  - Ruídos
    - Devidos a a problemas na aquisição
    - Ruídos de fontes externas
    - Imprecisão nas técnicas de medida
    - Erro de calibração
    - Erros de cálculos de dados derivados

Background

- O custo de dados imprecisos pode ser alto se eles forem utilizados para tomadas de decisões imediatas
- Dados imprecisos são uma das fontes de incertezas nas bases de dados.

Background

- Capacidades
  - Grande variação nas capacidades dos sensores e suas respectivas funcionalidades

Background

- Sensor Streaming x Streams tradicionais
  - Streams de sensores:
    - Normalmente amostragens, variadas, de dados
    - Imprecisa e com ruídos
    - Tamanho moderado
    - Alto custo de aquisição (nodos morrem)
  - Streams tradicionais
    - Conjunto completo de dados
    - Precisa
    - Tamanho elevado de dados
    - Baixo custo de aquisição

In-Network Storage

- Os dados são produzidos e precisam ser recuperados da rede para processamento e apresentação ao usuário
- A aquisição de dados deve ser energeticamente eficiente, escalável, auto organizável e robusto.
- Os sistemas atuais não respeitam estes princípios

In-Network Storage

- Novas formas para manipular os dados são necessárias.
  - Caches tradicionais não se aplicam pois não tem preocupações com as restrições de redes de sensores nem com correlações de tempo e espaço, comuns em redes de sensores.
  - *Geographic Information Systems* (GIS)
    - Processamento centralizado
    - Objetiva reduzir o custo de pesquisa
    - Compressão *Space first-time next*

In-Network Storage

- Três formas básicas de armazenar informações na rede
  - Externa
    - Os sensores produzem e enviam continuamente os dados para um ponto de acesso
  - Local
    - Os dados são armazenados localmente no sensor
  - Data centric
    - Os dados são nomeados e armazenados de acordo com este nome.
- Compromisso entre consumo de energia e forma de armazenamento

In-Network Storage

- Armazenamento externo
  - PREMON – PREdiction-based Monitoring
    - Tenta reduzir os custos de comunicação prevendo os futuros dados sensoriados. As predições são enviadas para os nodos que só retornam novos dados se estes estiverem fora de um *threshold* pré-definido
    - Problemas
      - Como generalizar a solução para pesquisas sumarizadas?
      - Aprender, online, padrões de *long* e *short term queries*.

In-Network Storage

- Armazenamento local
  - Dimensions
    - Habilita pesquisas em multi-resolução
    - Incorpora hierarquia e armazenamento distribuído
    - Utiliza-se de correlações espaço-temporais entre sensores e faz compressão local no tempo e distribui a compressão espacial para otimizar o consumo de energia (*wavelets*).
    - Problemas
      - Quantificar o ganho com compressão com relação a economia de energia

In-Network Storage

- Data centric
  - Geographical Hash Tables for Data Centric Storage
    - Observações freqüentes ou eventos
    - Dados massivos demais para saírem da rede
    - Os dados são nomeados e acessados através deste nome.
    - Todos os dados com o mesmo nome geral são armazenados no mesmo sensor
    - Caso haja uma pesquisa esta é enviada diretamente para este nodo, que pode ou não ter sido quem produziu o mesmo.

In-Network Storage

- Data centric
  - Geographical Hash Tables for Data Centric Storage
    - Problemas
      - As chaves do hash são distribuídas uniformemente no espaço geográfico. Caso os nodos não sejam distribuídos uniformemente, isto reduz a habilidade do algoritmo de distribuir as cargas de busca e armazenamento.
      - Nodos cientes de sua localização geográfica

In-Network Agregation

- A agregação pode ser distribuída ou centralizada
  - Para redes de sensores a forma centralizada tem um alto custo
  - Na agregação distribuída a *query* é feita e respondida pelos nodos da rede
- Roteamento e processamento dos dados não podem ser separados em redes de sensores

In-Network Agregation

- Apresentação de duas técnicas
  - TAG - Tiny AGgregation service for adhoc sensor networks
  - Aggregation for monitoring wireless sensor networks
  - Diferem basicamente no roteamento e em como a resposta a requisição é coletada

In-Network Agregation

- TAG - Tiny AGgregation service for adhoc sensor networks
  - Serviço de agregação genérico utilizando-se de uma linguagem SQL like para definição de requisições
  - Desenvolvido para aplicações remotas, difíceis de administrar
    - Monitoração de habitats, temperatura e utilização de energia
    - Somente dados sumarizados são necessários

In-Network Agregation

- Aggregation for monitoring wireless sensor networks
  - Tem como objetivo monitoramento da rede em busca de falhas e outras anomalias
  - Monitoramento contínuo de propriedades da rede

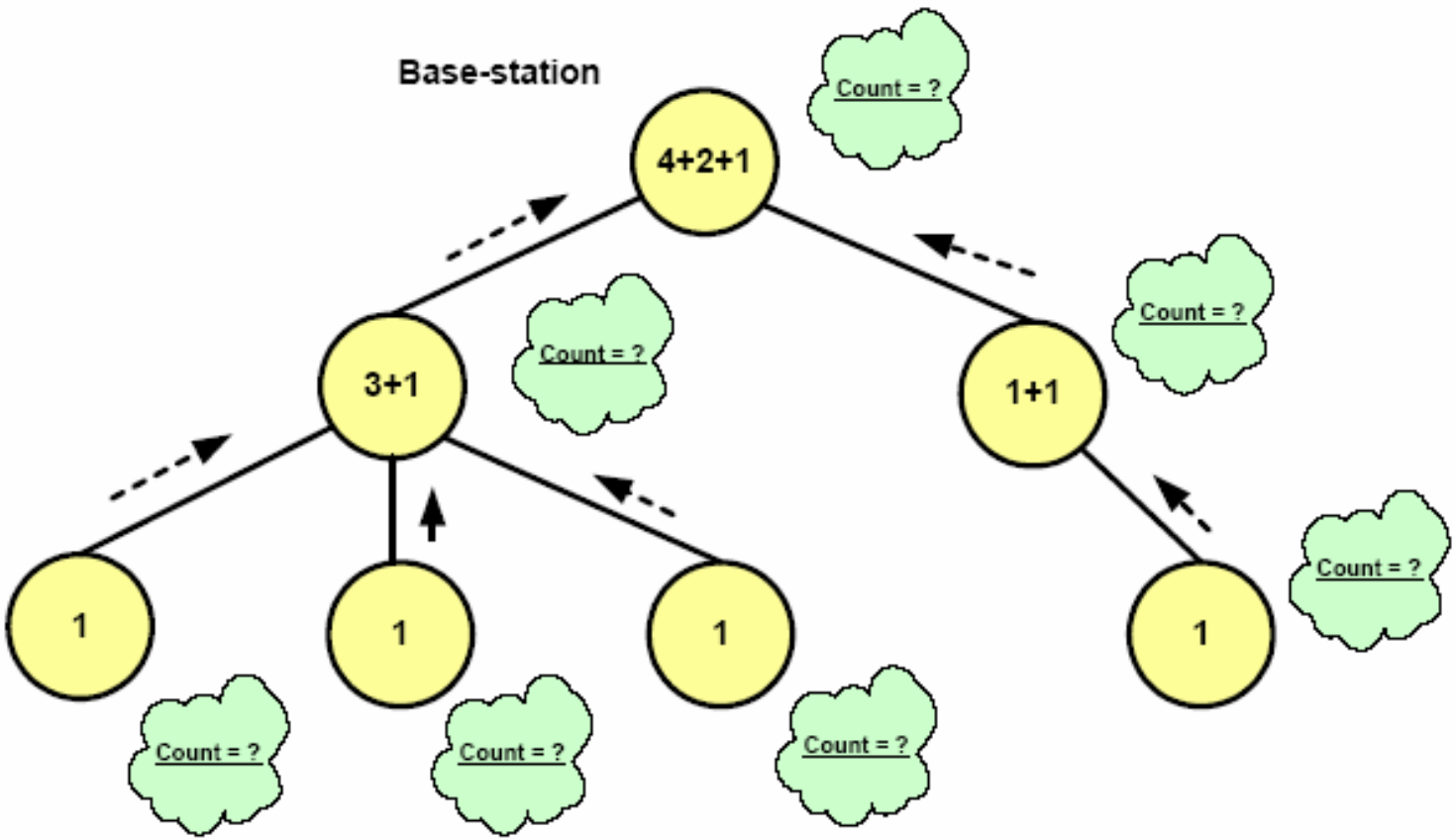
In-Network Agregation

- O modo básico de funcionamento das duas técnicas é similar
- As fases básicas são:
  - Avaliação da consulta
  - Roteamento

In-Network Agregation

- Avaliação da consulta
  - Composta de duas fases, distribuição e coleta
  - Distribuição
    - A consulta é distribuída para cada nó da rede
    - É montada uma árvore de roteamento tendo a ERB como raíz
    - Dados irrelevantes são descartados
    - Os dados relevantes são combinados de uma forma única e repassados para o pai.

In-Network Agregation



In-Network Agregation

- Avaliação da consulta
  - Coleta
    - Cada pai coleta os dados dos filhos em um intervalo pré-definido.
    - Eventualmente estes dados chegam na raiz.

In-Network Agregation

- Roteamento
  - TAG
    - Utiliza qualquer roteamento que proveja:
      - entrega da pesquisa para todos os nós da rede
      - Possa manter uma ou mais rotas de todos os nós até a raiz.
    - Utiliza-se de roteamento baseado em árvore onde a mensagem de construção da árvore tem o nível da raiz e cada nó adiciona um ao nível da mensagem que recebeu de seu pai e atualiza o pacote antes de retransmitir
    - Quando um nó quer enviar algo para a raiz envia para seu pai

In-Network Agregation

- Problemas
  - Indicado para monitoração de ambientes e onde os dados de interesse possam ser sumarizados, contudo de difícil adaptação para outros cenários
  - O histórico dos dados não é considerado
  - Overhead da construção da árvore
  - *Delay* na disponibilização dos dados ao cliente
  - O tempo de aquisição de dados escala linearmente com o tamanho da rede.
  - Sensível a perda de dados

Arquiteturas para Streams de Sensores

- Principal objetivo é recuperar os dados dos sensores de forma eficiente, do ponto de vista de energia
- Formas genéricas de modelagem de streams e de representação de redes de sensores como bases de dados

Arquiteturas para Streams de Sensores

- Sistemas e algoritmos tradicionais para bases de dados não são apropriados para streams em redes de sensores
  - Bases de dados tradicionais são baseadas em data-pull e execução off-line
  - Streams de dados de sensores são um fluxo massivo de dados periódicos que são “enviados” (push) continuamente para a base de dados
  - Buscas em redes de sensores têm uma componente temporal muito importante. As respostas se tornam inúteis se recebidas fora de seu prazo de validade
  - Mesmo algoritmos tradicionais para streams não funcionam em streams de redes de sensores pois normalmente não consideram as restrições de recursos

Arquiteturas para Streams de Sensores

- Fjording
  - A arquitetura genérica para buscas em streams de dados em redes de sensores
  - Três tipos de streams:
    - Históricas
      - Agregação de dados históricos
    - Snapshots
      - O valor sensoriado em um dado instante
    - Long running
      - Consultas contínuas em um dado intervalo de tempo

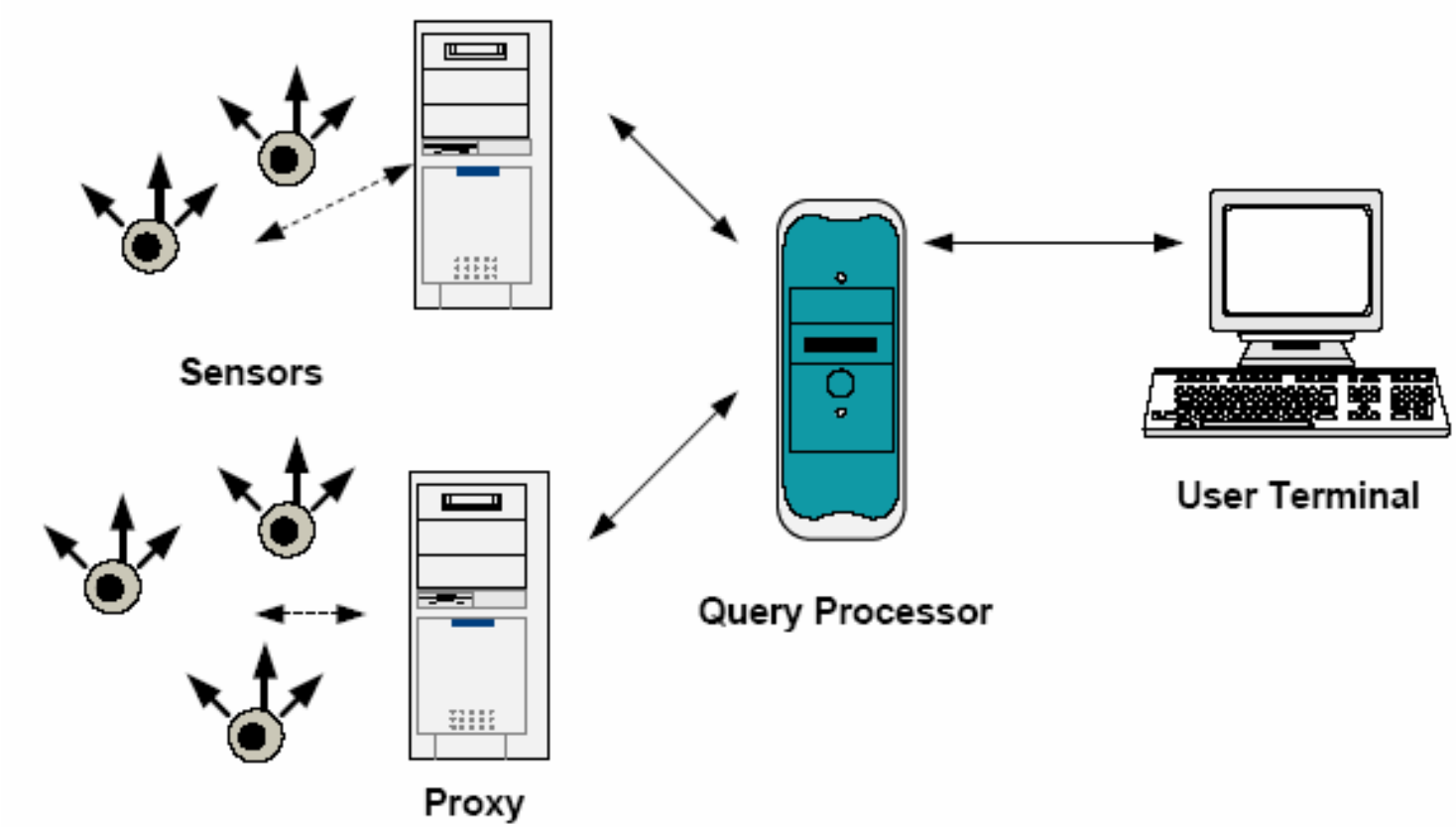
Arquiteturas para Streams de Sensores

- Fjording
  - Manipular múltiplas requisições tanto *push-based* quanto *pull-based*
  - Otimizar o uso dos recursos enquanto mantém um alto *throughput* de requisições
  - Dois elementos principais : fjord e sensor proxies

Arquiteturas para Streams de Sensores

- Fjording
  - O processador de requisições envia uma requisição para todos os proxies
  - Cada proxy verifica se tem sensores com dados relevantes associados a ele
  - Se tiver instrui estes a catalogar suas leituras
  - Os sensores enviam seus dados ao proxy, podendo os dados ser brutos ou processados
  - O proxy empacota os dados e envia para o servidor de requisições
  - O servidor de requisições processa a requisição e envia o resultado ao cliente

Arquiteturas para Streams de Sensores



Arquiteturas para Streams de Sensores

- Problemas a serem resolvidos em arquiteturas
  - Sincronização
  - *Delays* e falhas
  - Linguagens de requisição mais adequadas
  - Coleta de metadados sobre os sensores (posição, carga, capacidade)

Dados Errados

- Espera-se que os dados advindos do sensoriamento estejam imprecisos
- O grau de imprecisão varia com diversos fatores como custo do sensor, efeito do ambiente, etc.
- Os dados são normalmente provenientes de fenômenos físicos, estando portanto sujeitos a várias fontes de erro
- Pode ter alto custo se resultar na tomada imediata de decisão

Dados Errados

- Online Cleaning
  - Obter modelos “apurados” de incertezas
  - Modelos simples de predição nos sensores e modelos mais complexos no servidor de base de dados
  - Modelos de predição de erros devem ser associados a cada dado medido

Dados Errados

- Problema em aberto
  - Quais tarefas de “limpeza” dos dados são necessárias
  - Esta limpeza, devido as limitações dos sensores, pode ser feita de forma eficiente
  - Imprecisão multidimensional de dados
  - Dados adicionais podem ser utilizados para aumentar a confiança nos dados adquiridos, se sim quais dados são estes
  - Quais métricas usar

Conclusões

- Vários problemas e limitações das redes de sensores foram levantados e enfatizados
- Foi dado uma visão geral da pesquisa em armazenamento e recuperação de dados em redes de sensores e busca em streams
- Strams em redes de sensores são desafiadoras e diferentes de streams tradicionais
- Processamento local deve ser utilizado sempre que possível

Conclusões

- Os dados tem uma característica espaço-temporal que não pode ser ignorada
- Os ruídos podem interferir na busca de dados