Etapa 05 – Etapa Extração, Conversão e Melhoramento

MC536 – Bancos de Dados: Teoria e Prática Instituto de Computação Universidade Estadual de Campinas		2º semestre de 2015 Turmas A, B, C e D Professor: André Santanchè

Resumo

O objetivo desta etapa é preencher duas bases de dados especificada através da inclusão de informações a partir da extração de dados e do melhoramento de algumas destas informações.

Detalhamento

Os procedimentos de obtenção de dados da Web, para criação ou extensão de bancos de dados, estão associados a áreas como Extração de Informação, *Deep Web Mining*, e *Wrapper Induction*. No nosso caso, vamos utilizar dados da Web para preencher dois bancos de dados, um previamente especificado e outro especificado nesta etapa.

Experimento no Modelo de Grafos

A partir do modelo relacional produzido em etapas anteriores, as equipes devem produzir um modelo de grafos equivalente do banco, dando ênfase principalmente em como as relações devem ser exploradas para recomendação. Nesta etapa devem ser seguidos os passos:

1. Transformar o modelo relacional em grafo – atenção para os atributos do relacional que devem estar no grafo.
2. O modelo será apresentado na forma de exemplo, para isso deve ser usado o exemplo da etapa anterior, a ser convertido em grafo. Podem ser acrescentados novos elementos conforme a necessidade.
3. Escrever comandos em Cypher para implementar o exemplo. Pode ser usado o console (http://console.neo4j.org).
4. Fazer uma query de exemplo neste grafo.

O resultado deve ser submetido no mínimo por dupla e no máximo por metade do cluster (3 duplas) até o final do laboratório de hoje (18/09/2015).

Experimento no Modelo Relacional

O mesmo exemplo produzido nos grafos deve ser implementado no modelo relacional e devem ser construídas queries equivalentes no modelo relacional.

O resultado deve ser submetido pela mesma dupla ou sub-cluster (3 duplas) do experimento até o dia 24/09/2015.

Extração de Dados

As equipes devem obter os dados na Web e colocá-los nos bancos de dados relacional e de grafos. Há duas abordagens possíveis: (a) extrair simultaneamente para ambos; (b) extrair para a tabela, gerar um CSV e carregá-lo para o grafo.

Durante a extração serão percebidos novos requisitos para os esquemas. As equipes podem executar as transformações necessárias no esquema especificado para adequá-lo aos dados extraídos.

Limpeza de Dados (Data Cleaning)

Procedimentos de limpeza de dados são usados para remover redundâncias e melhorar a qualidade

dos dados armazenados. Na tabela relacional, os dados devem ser armazenados seguindo as formas normais apresentadas em classe. No modelo de grafos, deve haver cuidado à criação de índices.

Descoberta de Associações

Como este trabalho envolve a interligação de mais de uma base de dados, após a extração, os dados provenientes de bases diferentes devem ser interligados. Para isso, podem ser necessárias descobertas de associações entre dados usando: comparações aproximadas, outras bases que ajudem na associação etc.

Calendário de Submissões e Apresentação

- **18/09/2015** – cada dupla ou subcluster (máximo 3 duplas) submete um arquivo TXT contendo os comandos em Cypher do experimento criado em sala. O arquivo deve ter o nome apenas das pessoas presentes no lab. O nome do arquivo ZIP submetido deve seguir o padrão: cypher-duplasXX-YY-ZZ.txt em que XX, YY e ZZ são os números das duplas.
- **24/09/2015** – cada dupla ou subcluster (máximo 3 duplas) submete um arquivo TXT contendo os comandos em SQL do experimento. O arquivo deve ter o nome apenas das pessoas que participaram. O nome do arquivo ZIP submetido deve seguir o padrão: sql-duplasXX-YY-ZZ.txt em que XX, YY e ZZ são os números das duplas.
- **25/09 e 02/10/2015** – não haverá nenhuma apresentação formal, mas o professor, PEDs e PADs visitarão os cluster para saber o andamento dos trabalhos e cada cluster deve estar preparado para mostrar algo.
- **08/10/2015** - cada cluster deve submeter um arquivo ZIP contendo:
    - uma pasta chamada **[codigo]** contendo os códigos das aplicações que realizam as tarefas e instruções claras de compilação e execução;
    - uma pasta chamada **[dump]** com um arquivo de dump contendo os comandos SQL que geram o bando de dados da equipe até o momento ou uma parte dele contendo todos os esquemas mas um subconjunto dos registros;
    - uma pasta chamada **[slides]** com um arquivo PDF contendo os slides da apresentação.

  O nome do arquivo ZIP submetido deve seguir o padrão: extracao-clusterXX.zip
- **09/10/2015** - as equipes farão uma pequena apresentação no laboratório de seus resultados.

  Sugere-se que cada cluster prepare slides sumarizando aspectos importantes da extração de dados, limpeza de dados e descoberta de associações para apresentação em sala. Cada apresentação deve ter no máximo 20 minutos. Todos os membros da equipe devem estar presentes e estar preparados para participar da apresentação e responder perguntas.