Pesquisa e Publicação de Informação

Procura na Web

Nuno D. Mendes

Licenciatura em Sistemas e Tecnologias de Informação

25 Mai 2012

ISEGI – UNL

World Wide Web

- Repositório gigantesco ($\sim$1 Petabyte = 1000 Terabytes) de informação não-estruturada

World Wide Web

- Repositório gigantesco ($\sim$1 Petabyte = 1000 Terabytes) de informação não-estruturada
- Necessidade absoluta de desenvolvimento de sistemas de Pesquisa de Informação para procurar informação neste repositório: motores de pesquisa

World Wide Web

- Repositório gigantesco ($\sim$1 Petabyte = 1000 Terabytes) de informação não-estruturada
- Necessidade absoluta de desenvolvimento de sistemas de Pesquisa de Informação para procurar informação neste repositório: motores de pesquisa
- Desafios:

World Wide Web

- Repositório gigantesco ($\sim$1 Petabyte = 1000 Terabytes) de informação não-estruturada
- Necessidade absoluta de desenvolvimento de sistemas de Pesquisa de Informação para procurar informação neste repositório: motores de pesquisa
- Desafios:
  - **Dados distribuídos**: Os dados da web encontram-se em vários servidores e plataformas, nem sempre acessíveis e com larguras de banda variáveis.

World Wide Web

- Repositório gigantesco ($\sim$1 Petabyte = 1000 Terabytes) de informação não-estruturada
- Necessidade absoluta de desenvolvimento de sistemas de Pesquisa de Informação para procurar informação neste repositório: motores de pesquisa
- Desafios:
  - **Dados distribuídos**: Os dados da web encontram-se em vários servidores e plataformas, nem sempre acessíveis e com larguras de banda variáveis.
  - **Grande proporção de dados voláteis** : A web é um repositório extremamente dinâmico, com actualizações constantes, informação que é retirada ou deixa de estar acessível, e nova informação adicionada a cada segundo.

World Wide Web

- Repositório gigantesco ($\sim$1 Petabyte = 1000 Terabytes) de informação não-estruturada
- Necessidade absoluta de desenvolvimento de sistemas de Pesquisa de Informação para procurar informação neste repositório: motores de pesquisa
- Desafios:
  - **Dados distribuídos**: Os dados da web encontram-se em vários servidores e plataformas, nem sempre acessíveis e com larguras de banda variáveis.
  - **Grande proporção de dados voláteis** : A web é um repositório extremamente dinâmico, com actualizações constantes, informação que é retirada ou deixa de estar acessível, e nova informação adicionada a cada segundo.
  - **Grande dimensão do repositório** : A dimensão e crescimento exponencial da web dificultam grandemente a escalabilidade das soluções de Pesquisa de Informação adoptadas.

World Wide Web

- Repositório gigantesco ($\sim$1 Petabyte = 1000 Terabytes) de informação não-estruturada
- Necessidade absoluta de desenvolvimento de sistemas de Pesquisa de Informação para procurar informação neste repositório: motores de pesquisa
- Desafios:
  - **Dados distribuídos**: Os dados da web encontram-se em vários servidores e plataformas, nem sempre acessíveis e com larguras de banda variáveis.
  - **Grande proporção de dados voláteis** : A web é um repositório extremamente dinâmico, com actualizações constantes, informação que é retirada ou deixa de estar acessível, e nova informação adicionada a cada segundo.
  - **Grande dimensão do repositório** : A dimensão e crescimento exponencial da web dificultam grandemente a escalabilidade das soluções de Pesquisa de Informação adoptadas.
  - **Dados não-estruturados e redundantes** : Os dados da web não estão organizados, muitas páginas são cópias uma das outras, e há ainda uma grande proporção de redundância semântica.

World Wide Web

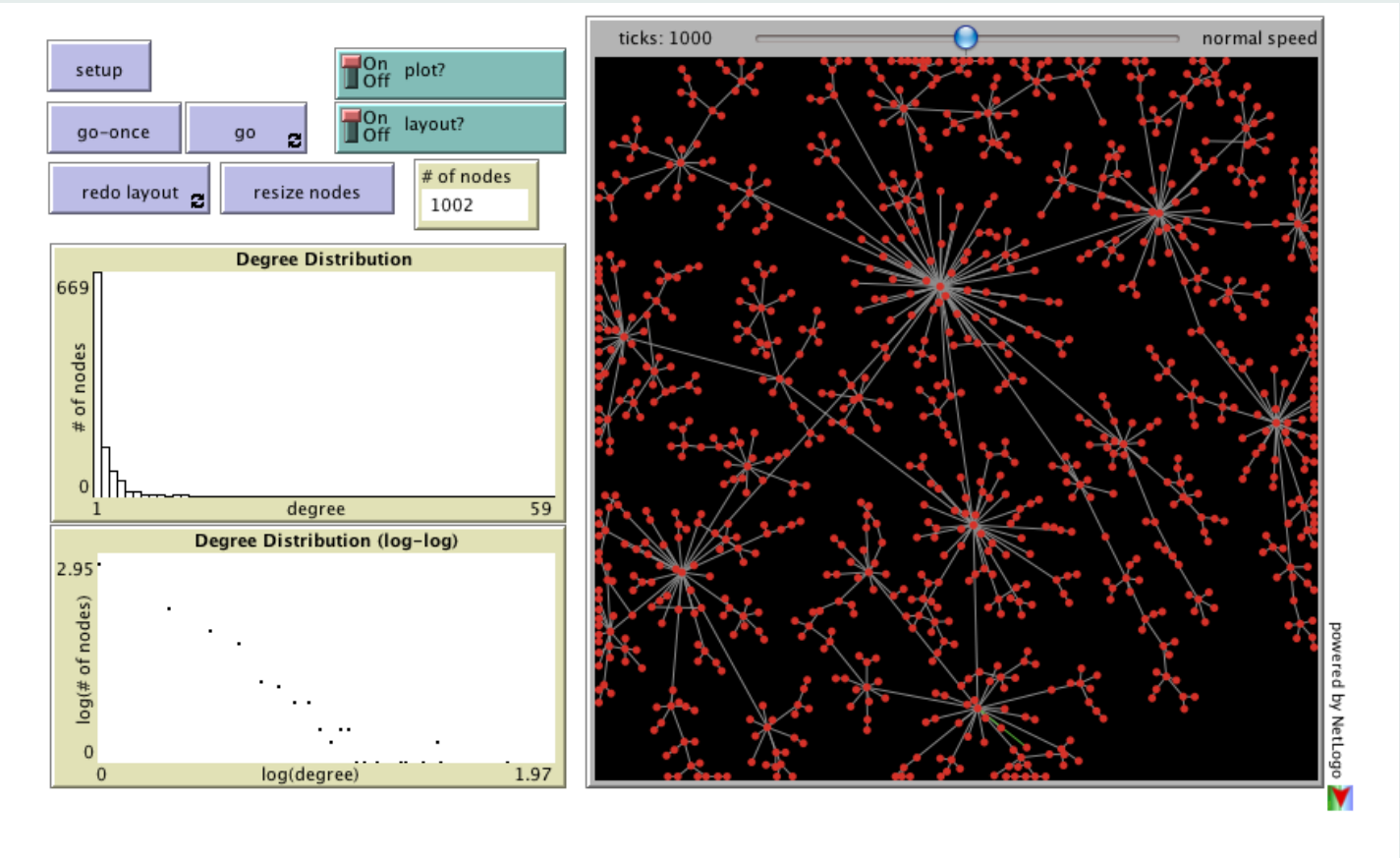
- Repositório gigantesco ($\sim$1 Petabyte = 1000 Terabytes) de informação não-estruturada
- Necessidade absoluta de desenvolvimento de sistemas de Pesquisa de Informação para procurar informação neste repositório: motores de pesquisa
- Desafios:
  - **Dados distribuídos**: Os dados da web encontram-se em vários servidores e plataformas, nem sempre acessíveis e com larguras de banda variáveis.
  - **Grande proporção de dados voláteis** : A web é um repositório extremamente dinâmico, com actualizações constantes, informação que é retirada ou deixa de estar acessível, e nova informação adicionada a cada segundo.
  - **Grande dimensão do repositório** : A dimensão e crescimento exponencial da web dificultam grandemente a escalabilidade das soluções de Pesquisa de Informação adoptadas.
  - **Dados não-estruturados e redundantes** : Os dados da web não estão organizados, muitas páginas são cópias uma das outras, e há ainda uma grande proporção de redundância semântica.
  - **Qualidade incerta dos dados** : Não existe um processo editorial na produção da dados da web. Muita informação pode ser falsa, incompleta, desactualizada, estar mal escrita ou conter erros tipográficos ou ortográficos.

World Wide Web

- Repositório gigantesco ($\sim$1 Petabyte = 1000 Terabytes) de informação não-estruturada
- Necessidade absoluta de desenvolvimento de sistemas de Pesquisa de Informação para procurar informação neste repositório: motores de pesquisa
- Desafios:
  - **Dados distribuídos**: Os dados da web encontram-se em vários servidores e plataformas, nem sempre acessíveis e com larguras de banda variáveis.
  - **Grande proporção de dados voláteis** : A web é um repositório extremamente dinâmico, com actualizações constantes, informação que é retirada ou deixa de estar acessível, e nova informação adicionada a cada segundo.
  - **Grande dimensão do repositório** : A dimensão e crescimento exponencial da web dificultam grandemente a escalabilidade das soluções de Pesquisa de Informação adoptadas.
  - **Dados não-estruturados e redundantes** : Os dados da web não estão organizados, muitas páginas são cópias uma das outras, e há ainda uma grande proporção de redundância semântica.
  - **Qualidade incerta dos dados** : Não existe um processo editorial na produção da dados da web. Muita informação pode ser falsa, incompleta, desactualizada, estar mal escrita ou conter erros tipográficos ou ortográficos.
  - **Dados heterogéneos** : Os dados da web, mesmo se nos restringirmos à informação textual, ocorrem em diversos formatos e em múltiplas línguas e alfabetos.

Caracterização da web

- O Google tem actualmente cerca de 50 000 000 000 de páginas indexadas (Maio 2012)
- Estima-se que existam cerca de 555 000 000 de websites (Dezembro 2011)



Tipo de distribuição de nós e suas interligações na Web

Motores de Pesquisa

- A web é vista como um repositório de texto
- Ao contrário de sistemas de PI comuns, o texto dos documentos pode não estar directamente disponível, mas apenas o índice
- Arquitecturas de motores de pesquisa:
  1. **Centralizada** : As páginas web são visitadas por um *Crawler* que vai preenchendo as entradas no *índice*. Uma *interface* recebe *queries* dos utilizadores que envia a um *motor de queries* que consulta usa o índice para encontrar páginas relevantes
  2. **Distribuída** : As tarefas de *crawling* e indexação são distribuídas por vários servidores, que se coordenam para determinar que páginas visitar e como combinar as ocorrências do índice distribuído

Relevância na web

- Ao contrário de outros repositórios de documentos, as páginas web encontram-se interligadas

Relevância na web

- Ao contrário de outros repositórios de documentos, as páginas web encontram-se interligadas
- O cálculo da relevância de um documento teve ter em conta esta característica particular

Relevância na web

- Ao contrário de outros repositórios de documentos, as páginas web encontram-se interligadas
- O cálculo da relevância de um documento teve ter em conta esta característica particular
- Um mecanismo famoso de determinação de relevância na web é o algoritmo `PageRank` do Google

Relevância na web

- Ao contrário de outros repositórios de documentos, as páginas web encontram-se interligadas
- O cálculo da relevância de um documento teve ter em conta esta característica particular
- Um mecanismo famoso de determinação de relevância na web é o algoritmo `PageRank` do Google

O `PageRank` simula a navegação aleatória na web. Partindo de uma página, a simulação salta para uma página aleatória com probabilidade q ou segue um dos links da página corrente com probabilidade $1 - q$. Admitimos que a simulação nunca regressa a uma página já visitada.
Seja $C(a)$ o número de links na página a, e sejam $p_1, \ldots, p_n$, páginas que contêm links para a. O *pagerank* de a é dado por:

$$\mathrm{PR}(a) = q + (1 - q) \sum_{i=1}^{n} \mathrm{PR}(p_i)/C(p_i)$$

Um valor típico para q é 0.15.

Crawlers

- Desafios:

Crawlers

- Desafios:
  1. Decidir que páginas visitar. Tipicamente parte-se de um conjunto de endereços e faz-se uma traversia em largura ou profundidade dos links de saída. Alternativamente, opta-se por ir visitando as páginas com maior estimativa de *rank*.

Crawlers

- Desafios:
  1. Decidir que páginas visitar. Tipicamente parte-se de um conjunto de endereços e faz-se uma traversia em largura ou profundidade dos links de saída. Alternativamente, opta-se por ir visitando as páginas com maior estimativa de *rank*.
  2. A visita exaustiva de páginas alojadas no mesmo servidor pode saturar a largura de banda ou a performance do servidor. Adoptam-se boas práticas para o comportamentos dos crawlers.

Crawlers

- Desafios:
  1. Decidir que páginas visitar. Tipicamente parte-se de um conjunto de endereços e faz-se uma traversia em largura ou profundidade dos links de saída. Alternativamente, opta-se por ir visitando as páginas com maior estimativa de *rank*.
  2. A visita exaustiva de páginas alojadas no mesmo servidor pode saturar a largura de banda ou a performance do servidor. Adoptam-se boas práticas para o comportamentos dos crawlers.
  3. É necessário guardar a data de visita das páginas para agendar uma nova visita ou estimar se a informação está actualizada ou desactualizada no índice.

Crawlers

- Desafios:
  1. Decidir que páginas visitar. Tipicamente parte-se de um conjunto de endereços e faz-se uma traversia em largura ou profundidade dos links de saída. Alternativamente, opta-se por ir visitando as páginas com maior estimativa de *rank*.
  2. A visita exaustiva de páginas alojadas no mesmo servidor pode saturar a largura de banda ou a performance do servidor. Adoptam-se boas práticas para o comportamentos dos crawlers.
  3. É necessário guardar a data de visita das páginas para agendar uma nova visita ou estimar se a informação está actualizada ou desactualizada no índice.
  4. Páginas visitadas e indexadas podem entretanto ter desaparecido ou já não serem acessíveis.

Crawlers

- Desafios:
  1. Decidir que páginas visitar. Tipicamente parte-se de um conjunto de endereços e faz-se uma traversia em largura ou profundidade dos links de saída. Alternativamente, opta-se por ir visitando as páginas com maior estimativa de *rank*.
  2. A visita exaustiva de páginas alojadas no mesmo servidor pode saturar a largura de banda ou a performance do servidor. Adoptam-se boas práticas para o comportamentos dos crawlers.
  3. É necessário guardar a data de visita das páginas para agendar uma nova visita ou estimar se a informação está actualizada ou desactualizada no índice.
  4. Páginas visitadas e indexadas podem entretanto ter desaparecido ou já não serem acessíveis.
  5. Páginas órfãs dificilmente são encontradas.

Crawlers

- Desafios:
  1. Decidir que páginas visitar. Tipicamente parte-se de um conjunto de endereços e faz-se uma traversia em largura ou profundidade dos links de saída. Alternativamente, opta-se por ir visitando as páginas com maior estimativa de *rank*.
  2. A visita exaustiva de páginas alojadas no mesmo servidor pode saturar a largura de banda ou a performance do servidor. Adoptam-se boas práticas para o comportamentos dos crawlers.
  3. É necessário guardar a data de visita das páginas para agendar uma nova visita ou estimar se a informação está actualizada ou desactualizada no índice.
  4. Páginas visitadas e indexadas podem entretanto ter desaparecido ou já não serem acessíveis.
  5. Páginas órfãs dificilmente são encontradas.
  6. Páginas em formatos não-HTML necessitam de tratamento especial.

Indexação

- Tipicamente são usados ficheiros invertidos pela simplicidade e economia do tamanho do índice.

Indexação

- Tipicamente são usados ficheiros invertidos pela simplicidade e economia do tamanho do índice.
- A granularidade do índice é comummente a página, mas alguns motores de pesquisa indexam as posições dos termos

Indexação

- Tipicamente são usados ficheiros invertidos pela simplicidade e economia do tamanho do índice.
- A granularidade do índice é comummente a página, mas alguns motores de pesquisa indexam as posições dos termos
- Vários motores de pesquisa actualmente suportam pesquisa de frases ou termos compostos. Os detalhes de implementação não são públicos.

Indexação

- Tipicamente são usados ficheiros invertidos pela simplicidade e economia do tamanho do índice.
- A granularidade do índice é comummente a página, mas alguns motores de pesquisa indexam as posições dos termos
- Vários motores de pesquisa actualmente suportam pesquisa de frases ou termos compostos. Os detalhes de implementação não são públicos.

Personalização

Certos motores de pesquisa incorporam o histórico de pesquisas e o idioma preferencial do utilizador na determinação da relevância dos documentos, numa tentativa de melhor antecipar as necessidades de informação de cada utilizador.