Organização e Recuperação de Informação: Análise de links

Marcelo K. Albertini

Faculdade de Computação, Universidade Federal de Uberlândia

Conteúdo

- Texto de âncora: o que são links e porquê são importantes para ORI

Conteúdo

- Texto de âncora: o que são links e porquê são importantes para ORI

Conteúdo

- Texto de âncora: o que são links e porquê são importantes para ORI
- Análise de citação: fundação matemática de PageRank e ranking baseado em links

Conteúdo

- Texto de âncora: o que são links e porquê são importantes para ORI
- Análise de citação: fundação matemática de PageRank e ranking baseado em links
- PageRank: algoritmo original

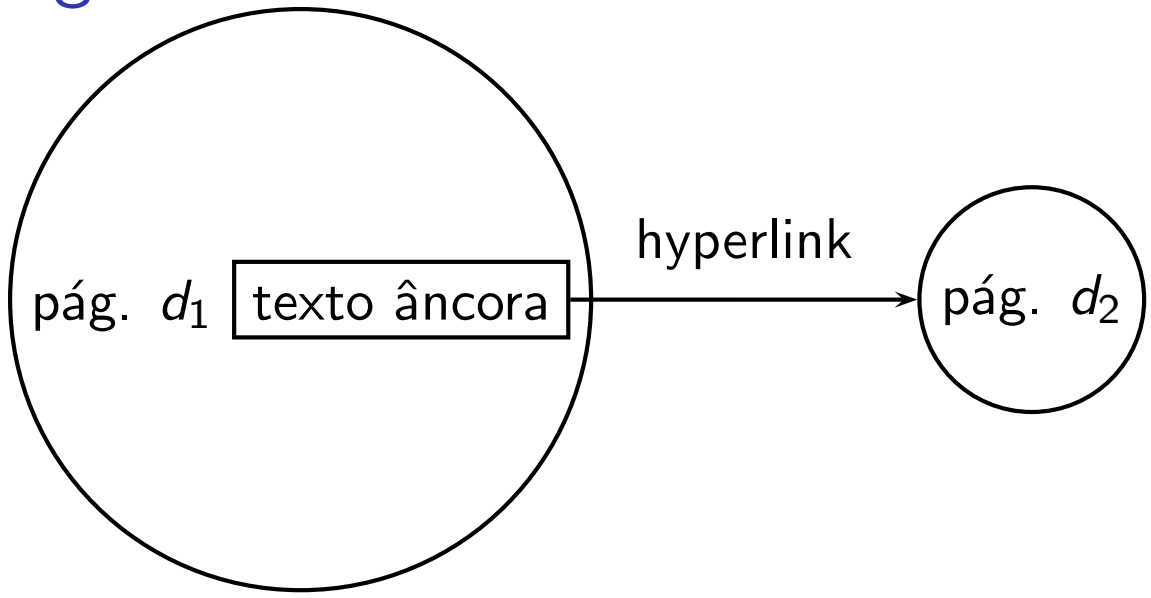
Conteúdo

- Texto de âncora: o que são links e porquê são importantes para ORI
- Análise de citação: fundação matemática de PageRank e ranking baseado em links
- PageRank: algoritmo original
- Centros & Autoridades: um algoritmo alternativo de ranking baseados em links

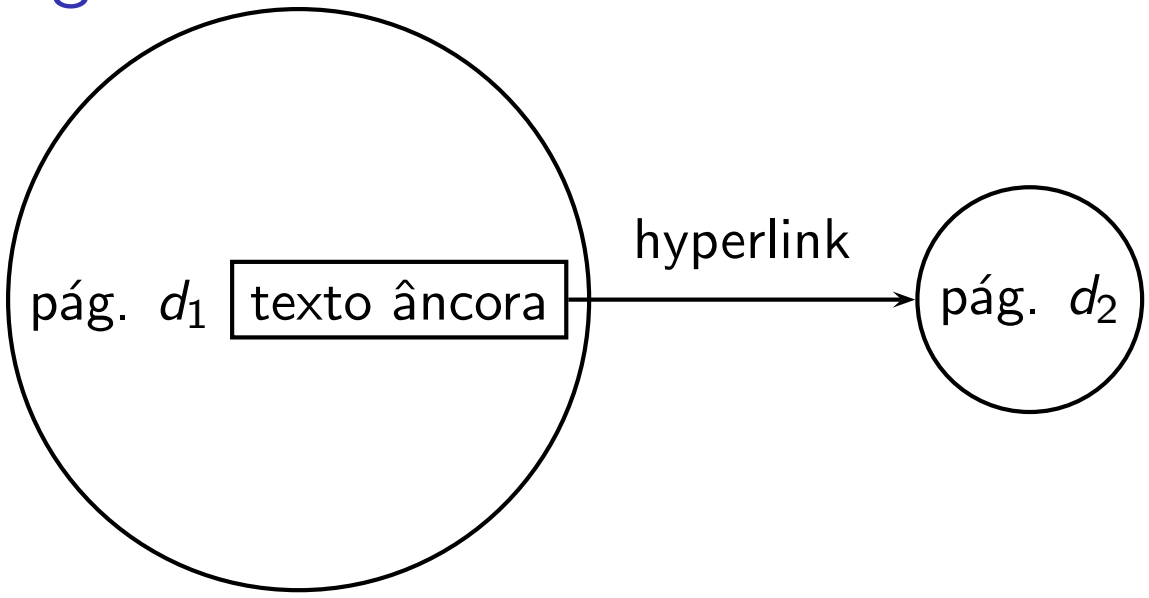
Conteúdo

A web: um grafo direcionado

A web: um grafo direcionado

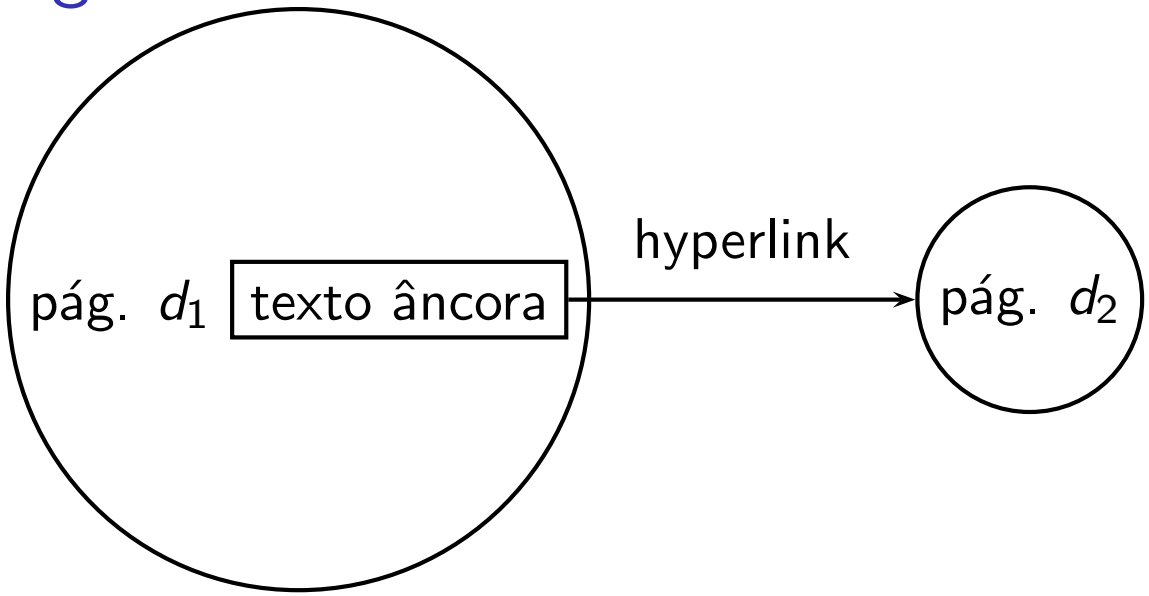


A web: um grafo direcionado



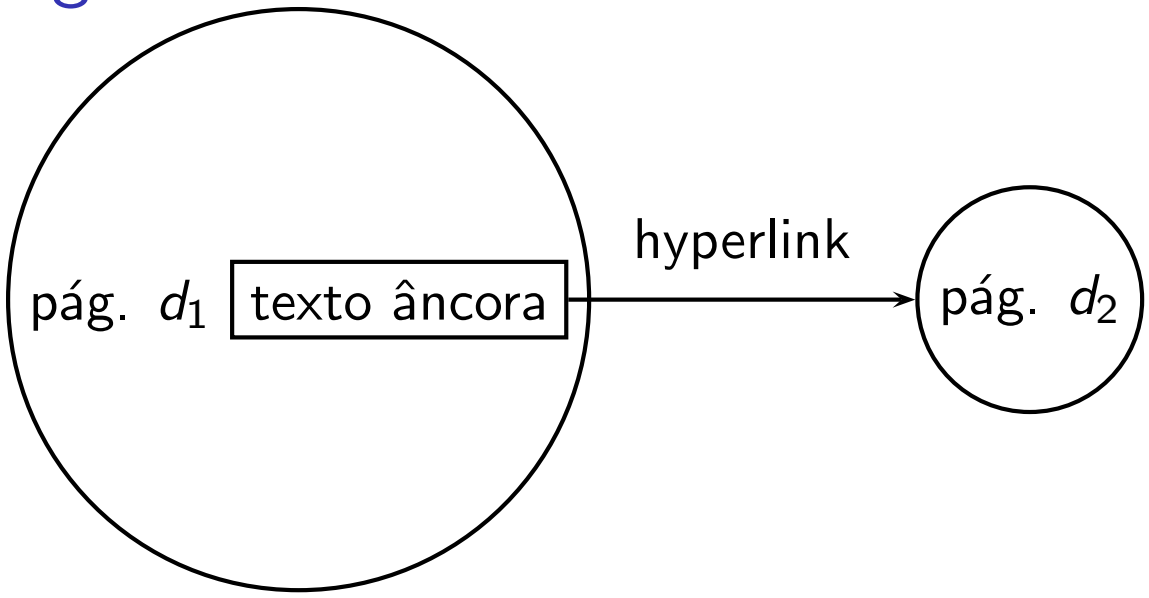
- Premissa 1: Um hyperlink é um sinal de qualidade.

A web: um grafo direcionado



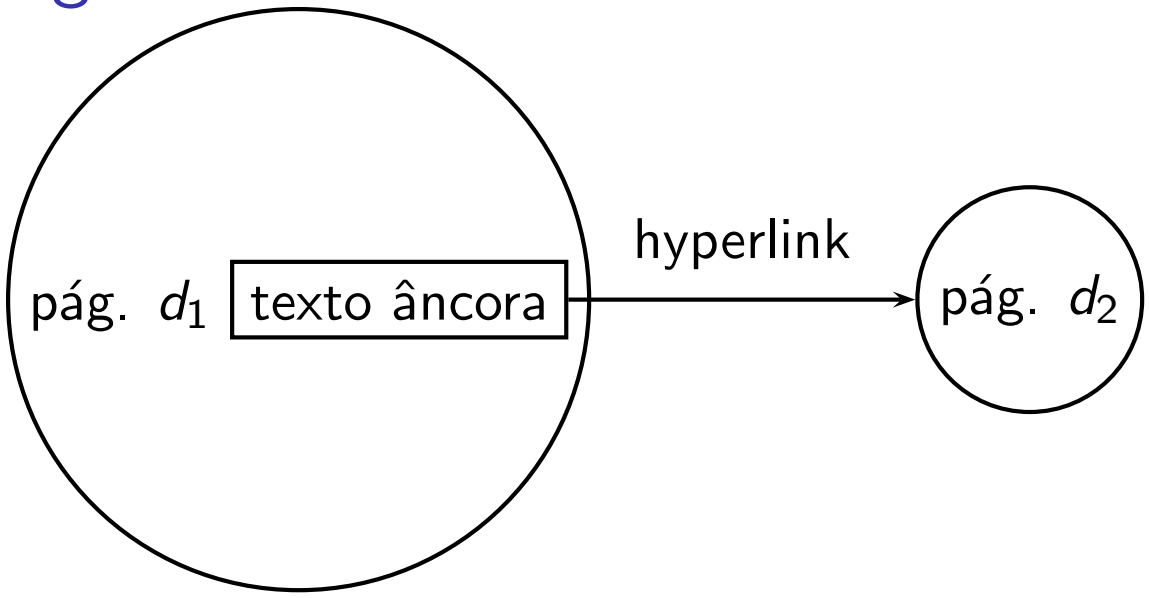
- Premissa 1: Um hyperlink é um sinal de qualidade.
  - O hyperlink $d_1 \rightarrow d_2$ indica que o autor de d_1 considera d_2 como sendo boa qualidade e relevante.

A web: um grafo direcionado



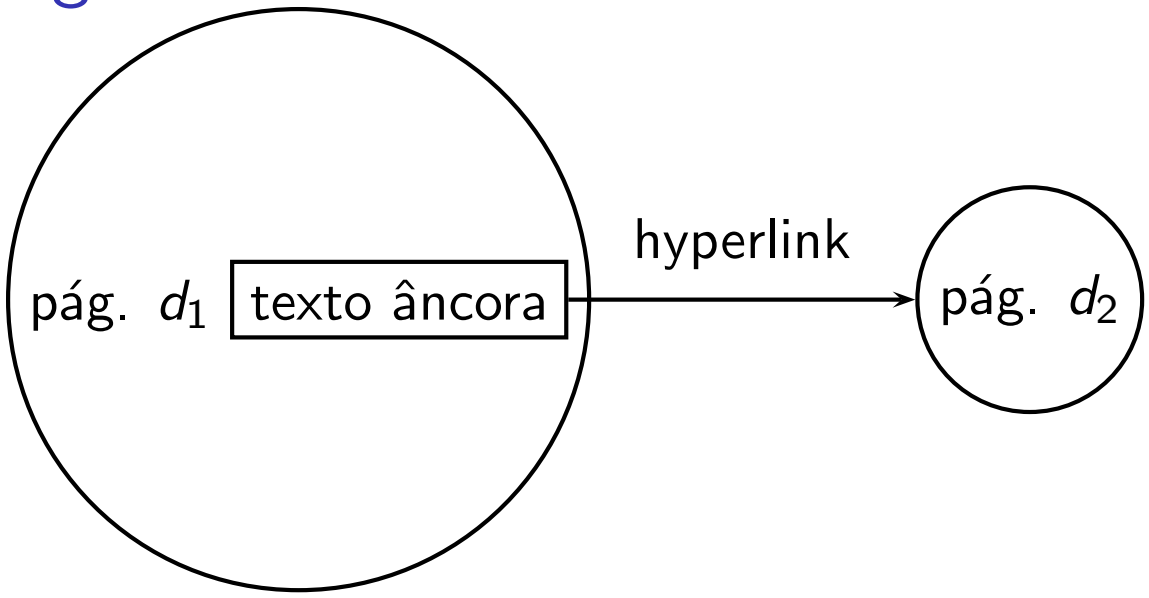
- Premissa 1: Um hyperlink é um sinal de qualidade.
  - O hyperlink $d_1 \rightarrow d_2$ indica que o autor de d_1 considera d_2 como sendo boa qualidade e relevante.
- Premissa 2: O texto âncora descreve o conteúdo d_2.

A web: um grafo direcionado



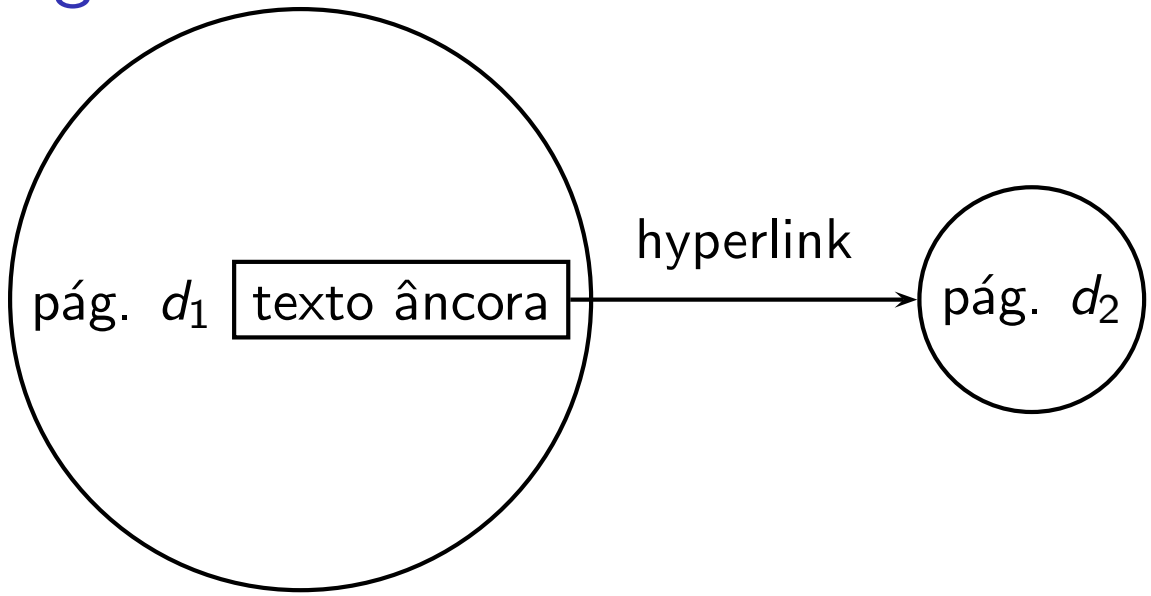
- Premissa 1: Um hyperlink é um sinal de qualidade.
  - O hyperlink $d_1 \rightarrow d_2$ indica que o autor de d_1 considera d_2 como sendo boa qualidade e relevante.
- Premissa 2: O texto âncora descreve o conteúdo d_2.
  - Consideramos o texto âncora sendo o texto em volta do hyperlink.

A web: um grafo direcionado



- Premissa 1: Um hyperlink é um sinal de qualidade.
  - O hyperlink $d_1 \rightarrow d_2$ indica que o autor de d_1 considera d_2 como sendo boa qualidade e relevante.
- Premissa 2: O texto âncora descreve o conteúdo d_2.
  - Consideramos o texto âncora sendo o texto em volta do hyperlink.
  - Exemplo: "Encontre carros baratos aqui."

A web: um grafo direcionado



- Premissa 1: Um hyperlink é um sinal de qualidade.
  - O hyperlink $d_1 \rightarrow d_2$ indica que o autor de d_1 considera d_2 como sendo boa qualidade e relevante.
- Premissa 2: O texto âncora descreve o conteúdo d_2.
  - Consideramos o texto âncora sendo o texto em volta do hyperlink.
  - Exemplo: "Encontre carros baratos aqui."
  - texto âncora: "Encontre carros baratos aqui"
  - Usando a definição formal: apenas texto visível em um hyperlink: aqui

[texto de d_2] somente vs. [texto de d_2] + [texto âncora $\rightarrow d_2$]

- Buscar no [texto de d_2] + [texto âncora $\rightarrow d_2$] costuma ser mais efetivo que buscar no [texto de d_2] somente

[texto de d_2] somente vs. [texto de d_2] + [texto âncora $\rightarrow d_2$]

- Buscar no [texto de d_2] + [texto âncora $\rightarrow d_2$] costuma ser mais efetivo que buscar no [texto de d_2] somente
- Exemplo: consulta: *IBM*

[texto de d_2] somente vs. [texto de d_2] + [texto âncora $\rightarrow d_2$]

- Buscar no [texto de d_2] + [texto âncora $\rightarrow d_2$] costuma ser mais efetivo que buscar no [texto de d_2] somente
- Exemplo: consulta: *IBM*
  - Retorna página de direitos autorais da IBM

[texto de d_2] somente vs. [texto de d_2] + [texto âncora $\rightarrow d_2$]

- Buscar no [texto de d_2] + [texto âncora $\rightarrow d_2$] costuma ser mais efetivo que buscar no [texto de d_2] somente
- Exemplo: consulta: *IBM*
  - Retorna página de direitos autorais da IBM
  - Retorna muitas páginas de spam

[texto de d_2] somente vs. [texto de d_2] + [texto âncora $\rightarrow d_2$]

- Buscar no [texto de d_2] + [texto âncora $\rightarrow d_2$] costuma ser mais efetivo que buscar no [texto de d_2] somente
- Exemplo: consulta: *IBM*
  - Retorna página de direitos autorais da IBM
  - Retorna muitas páginas de spam
  - Retorna artigo do Wikipedia sobre a IBM

[texto de d_2] somente vs. [texto de d_2] + [texto âncora $\rightarrow d_2$]

- Buscar no [texto de d_2] + [texto âncora $\rightarrow d_2$] costuma ser mais efetivo que buscar no [texto de d_2] somente
- Exemplo: consulta: *IBM*
  - Retorna página de direitos autorais da IBM
  - Retorna muitas páginas de spam
  - Retorna artigo do Wikipedia sobre a IBM
  - Mas talvez não retorne a home page da IBM

[texto de d_2] somente vs. [texto de d_2] + [texto âncora $\rightarrow d_2$]

- Buscar no [texto de d_2] + [texto âncora $\rightarrow d_2$] costuma ser mais efetivo que buscar no [texto de d_2] somente
- Exemplo: consulta: *IBM*
  - Retorna página de direitos autorais da IBM
  - Retorna muitas páginas de spam
  - Retorna artigo do Wikipedia sobre a IBM
  - Mas talvez não retorne a home page da IBM
  - . . . se a home page for a maior parte imagens

[texto de d_2] somente vs. [texto de d_2] + [texto âncora $\rightarrow d_2$]

- Buscar no [texto de d_2] + [texto âncora $\rightarrow d_2$] costuma ser mais efetivo que buscar no [texto de d_2] somente
- Exemplo: consulta: *IBM*
  - Retorna página de direitos autorais da IBM
  - Retorna muitas páginas de spam
  - Retorna artigo do Wikipedia sobre a IBM
  - Mas talvez não retorne a home page da IBM
  - . . . se a home page for a maior parte imagens
- Buscar no [texto âncora $\rightarrow d_2$] é melhor para a consulta *IBM*.

[texto de d_2] somente vs. [texto de d_2] + [texto âncora $\rightarrow d_2$]

- Buscar no [texto de d_2] + [texto âncora $\rightarrow d_2$] costuma ser mais efetivo que buscar no [texto de d_2] somente
- Exemplo: consulta: *IBM*
  - Retorna página de direitos autorais da IBM
  - Retorna muitas páginas de spam
  - Retorna artigo do Wikipedia sobre a IBM
  - Mas talvez não retorne a home page da IBM
  - . . . se a home page for a maior parte imagens
- Buscar no [texto âncora $\rightarrow d_2$] é melhor para a consulta *IBM*.
  - Nessa representação, a página com maior número de ocorrências de *IBM* é www.ibm.com.

Texto âncora contendo *IBM* apontando para www.ibm.com

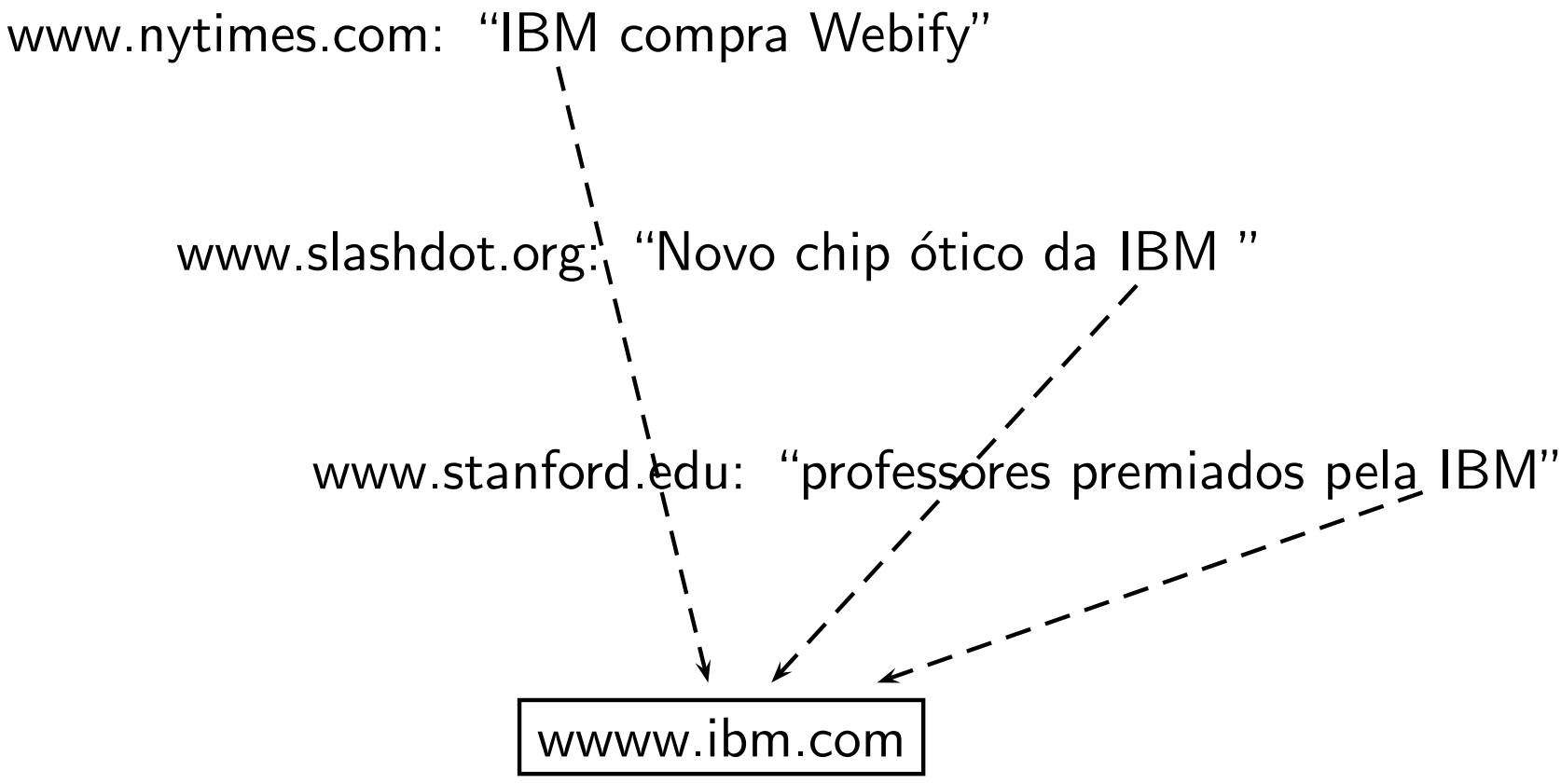
www.nytimes.com: "IBM compra Webify"

www.slashdot.org: "Novo chip ótico da IBM "

www.stanford.edu: "professores premiados pela IBM"

wwww.ibm.com

Texto âncora contendo *IBM* apontando para www.ibm.com



Indexação de texto âncora

- Portanto, texto âncora pode ser uma descrição melhor da página, que o próprio conteúdo da página

Indexação de texto âncora

- Portanto, texto âncora pode ser uma descrição melhor da página, que o próprio conteúdo da página
- texto âncora pode ter atribuído um peso mais alto que o texto dos documentos. (baseado nas Premissas 1&2)

Indexação de texto âncora

- Portanto, texto âncora pode ser uma descrição melhor da página, que o próprio conteúdo da página
- texto âncora pode ter atribuído um peso mais alto que o texto dos documentos. (baseado nas Premissas 1&2)
- Indexação de texto âncora pode ter efeitos colaterais indesejáveis: bombardeio de links

Premissas do PageRank

- Premissa 1: Um link na web é um sinal de qualidade – o autor do link pensa que a página referenciada é de alta qualidade

Premissas do PageRank

- Premissa 1: Um link na web é um sinal de qualidade – o autor do link pensa que a página referenciada é de alta qualidade
- Premissa 2: O texto âncora descreve o conteúdo da página referenciada

Premissas do PageRank

- Premissa 1: Um link na web é um sinal de qualidade – o autor do link pensa que a página referenciada é de alta qualidade
- Premissa 2: O texto âncora descreve o conteúdo da página referenciada
- A premissa 1 é verdadeira em geral?

Premissas do PageRank

- Premissa 1: Um link na web é um sinal de qualidade – o autor do link pensa que a página referenciada é de alta qualidade
- Premissa 2: O texto âncora descreve o conteúdo da página referenciada
- A premissa 1 é verdadeira em geral?
- A premissa 2 é verdadeira em geral?

Bombardeio de links

- Um bombardeio de links é uma busca maliciosamente manipulada a partir do texto âncora.

Bombardeio de links

- Um bombardeio de links é uma busca maliciosamente manipulada a partir do texto âncora.
- `déspota cachaceiro`, `more evil than Satan`

Bombardeio de links

- Um bombardeio de links é uma busca maliciosamente manipulada a partir do texto âncora.
- `déspota cachaceiro`, `more evil than Satan`
- Em 2007, novas função de pesos para o ranking baseado em links corrigiu muitos bombardeio de links

Bombardeio de links

- Um bombardeio de links é uma busca maliciosamente manipulada a partir do texto âncora.
- `déspota cachaceiro`, `more evil than Satan`
- Em 2007, novas função de pesos para o ranking baseado em links corrigiu muitos bombardeio de links
- Ainda existem alguns: [dangerous cult] no DuckDuckGo (7º), Google (3º), Bing (9º), Yahoo (9º),

Bombardeio de links

- Um bombardeio de links é uma busca maliciosamente manipulada a partir do texto âncora.
- `déspota cachaceiro`, `more evil than Satan`
- Em 2007, novas função de pesos para o ranking baseado em links corrigiu muitos bombardeio de links
- Ainda existem alguns: [dangerous cult] no DuckDuckGo (7º), Google (3º), Bing (9º), Yahoo (9º),
  - Criação coordenada de links por parte de pessoas que não gostam da Igreja da Cientologia

Conteúdo

Origem do PageRank: Análise de citação (1)

- Análise de citação: análise de citações na literatura científica

Origem do PageRank: Análise de citação (1)

- Análise de citação: análise de citações na literatura científica
- Exemplo de citação: "Miller (2001) demonstrou que atividade física altera o metabolismo de estrogênio."

Origem do PageRank: Análise de citação (1)

- Análise de citação: análise de citações na literatura científica
- Exemplo de citação: "Miller (2001) demonstrou que atividade física altera o metabolismo de estrogênio."
- Podemos considerar "Miller (2001)" como um link entre dois artigos científicos

Origem do PageRank: Análise de citação (1)

- Análise de citação: análise de citações na literatura científica
- Exemplo de citação: "Miller (2001) demonstrou que atividade física altera o metabolismo de estrogênio."
- Podemos considerar "Miller (2001)" como um link entre dois artigos científicos
- Aplicação desses "hyperlinks" na literatura científica:

Origem do PageRank: Análise de citação (1)

- Análise de citação: análise de citações na literatura científica
- Exemplo de citação: "Miller (2001) demonstrou que atividade física altera o metabolismo de estrogênio."
- Podemos considerar "Miller (2001)" como um link entre dois artigos científicos
- Aplicação desses "hyperlinks" na literatura científica:
  - Medir a similaridade de dois artigos pelos artigos citados por ambos ou artigos em comuns citando ambos

Origem do PageRank: Análise de citação (1)

- Análise de citação: análise de citações na literatura científica
- Exemplo de citação: "Miller (2001) demonstrou que atividade física altera o metabolismo de estrogênio."
- Podemos considerar "Miller (2001)" como um link entre dois artigos científicos
- Aplicação desses "hyperlinks" na literatura científica:
  - Medir a similaridade de dois artigos pelos artigos citados por ambos ou artigos em comuns citando ambos
  - Similaridade de co-citação

Origem do PageRank: Análise de citação (1)

- Análise de citação: análise de citações na literatura científica
- Exemplo de citação: "Miller (2001) demonstrou que atividade física altera o metabolismo de estrogênio."
- Podemos considerar "Miller (2001)" como um link entre dois artigos científicos
- Aplicação desses "hyperlinks" na literatura científica:
  - Medir a similaridade de dois artigos pelos artigos citados por ambos ou artigos em comuns citando ambos
  - Similaridade de co-citação
  - Similaridade de co-citação na web: operador "related:" do Google , e.g. [related:www.ford.com]

Origem do PageRank: Análise de citação (2)

- Aplicação: frequência de citação pode ser usada para medir o impacto de um artigo científico

Origem do PageRank: Análise de citação (2)

- Aplicação: frequência de citação pode ser usada para medir o impacto de um artigo científico
  - Medida mais simples: cada citação recebe um voto – não é muito preciso

Origem do PageRank: Análise de citação (2)

- Aplicação: frequência de citação pode ser usada para medir o impacto de um artigo científico
  - Medida mais simples: cada citação recebe um voto – não é muito preciso
- Na web: frequência de citação = contagem inlink = links recebidos

Origem do PageRank: Análise de citação (2)

- Aplicação: frequência de citação pode ser usada para medir o impacto de um artigo científico
  - Medida mais simples: cada citação recebe um voto – não é muito preciso
- Na web: frequência de citação = contagem inlink = links recebidos
  - Uma contagem de inlink alta não reflete em alta qualidade ...

Origem do PageRank: Análise de citação (2)

- Aplicação: frequência de citação pode ser usada para medir o impacto de um artigo científico
  - Medida mais simples: cada citação recebe um voto – não é muito preciso
- Na web: frequência de citação = contagem inlink = links recebidos
  - Uma contagem de inlink alta não reflete em alta qualidade ...
  - ... por causa principalmente de spam

Origem do PageRank: Análise de citação (2)

- Aplicação: frequência de citação pode ser usada para medir o impacto de um artigo científico
  - Medida mais simples: cada citação recebe um voto – não é muito preciso
- Na web: frequência de citação = contagem inlink = links recebidos
  - Uma contagem de inlink alta não reflete em alta qualidade ...
  - ... por causa principalmente de spam
- Medida melhor: frequência de citação ponderada ou rank de citações

Origem do PageRank: Análise de citação (2)

- Aplicação: frequência de citação pode ser usada para medir o impacto de um artigo científico
  - Medida mais simples: cada citação recebe um voto – não é muito preciso
- Na web: frequência de citação = contagem inlink = links recebidos
  - Uma contagem de inlink alta não reflete em alta qualidade ...
  - ... por causa principalmente de spam
- Medida melhor: frequência de citação ponderada ou rank de citações
  - Um voto de citação é ponderado de acordo com o seu impacto de citação

Origem do PageRank: Análise de citação (2)

- Aplicação: frequência de citação pode ser usada para medir o impacto de um artigo científico
  - Medida mais simples: cada citação recebe um voto – não é muito preciso
- Na web: frequência de citação = contagem inlink = links recebidos
  - Uma contagem de inlink alta não reflete em alta qualidade ...
  - ... por causa principalmente de spam
- Medida melhor: frequência de citação ponderada ou rank de citações
  - Um voto de citação é ponderado de acordo com o seu impacto de citação
  - Circular? Não: pode ser formalizado em um modo bem definido

Origem do PageRank: Análise de citação (3)

- Medida melhor: frequência de citação ponderada ou rank de citações

Origem do PageRank: Análise de citação (3)

- Medida melhor: frequência de citação ponderada ou rank de citações
- Isto é, basicamente, o PageRank.

Origem do PageRank: Análise de citação (3)

- Medida melhor: frequência de citação ponderada ou rank de citações
- Isto é, basicamente, o PageRank.
- PageRank foi inventado no contexto de análise de citações por Pinsker e Narin nos anos 1960s.

Origem do PageRank: Análise de citação (3)

- Medida melhor: frequência de citação ponderada ou rank de citações
- Isto é, basicamente, o PageRank.
- PageRank foi inventado no contexto de análise de citações por Pinsker e Narin nos anos 1960s.
- Análise de citação é importante: o salário de professores, investimento em projetos de pesquisa, infraestrutura da universidade são definidos pelo impacto da pesquisa avaliado pela análise de citações

Origem do PageRank: resumo

- Podemos usar o mesmo modelo para

Origem do PageRank: resumo

- Podemos usar o mesmo modelo para
  - citações em literatura científica

Origem do PageRank: resumo

- Podemos usar o mesmo modelo para
  - citações em literatura científica
  - hyperlinks na web

Origem do PageRank: resumo

- Podemos usar o mesmo modelo para
  - citações em literatura científica
  - hyperlinks na web
- frequência de citação com ponderação adequada é uma medida excelente de qualidade . . .

Origem do PageRank: resumo

- Podemos usar o mesmo modelo para
  - citações em literatura científica
  - hyperlinks na web
- frequência de citação com ponderação adequada é uma medida excelente de qualidade . . .
  - . . . tanto para páginas web quanto para publicações científicas

Origem do PageRank: resumo

- Podemos usar o mesmo modelo para
  - citações em literatura científica
  - hyperlinks na web
- frequência de citação com ponderação adequada é uma medida excelente de qualidade . . .
  - . . . tanto para páginas web quanto para publicações científicas
- Próximo: algoritmo PageRank para calcular a frequência de citações ponderada para a web

Ranking baseado em links para busca web

Ranking baseado em links para busca web

- Versão simples usando links para ranking na web

Ranking baseado em links para busca web

- Versão simples usando links para ranking na web
  - Primeiro: capturar todas as páginas adequadas para a consulta

Ranking baseado em links para busca web

- Versão simples usando links para ranking na web
  - Primeiro: capturar todas as páginas adequadas para a consulta
  - Ordenar as páginas pelo número de inlinks (links recebidos)

Ranking baseado em links para busca web

- Versão simples usando links para ranking na web
  - Primeiro: capturar todas as páginas adequadas para a consulta
  - Ordenar as páginas pelo número de inlinks (links recebidos)
- Usando somente popularidade de links é fácil de fazer spam. Porquê?

Conteúdo

Modelo do PageRank: Caminhada aleatória

- Imagine um navegador fazendo caminhada aleatória na web

Modelo do PageRank: Caminhada aleatória

- Imagine um navegador fazendo caminhada aleatória na web
  - Iniciar em uma página aleatória

Modelo do PageRank: Caminhada aleatória

- Imagine um navegador fazendo caminhada aleatória na web
  - Iniciar em uma página aleatória
  - A cada passo, sair da página atual e ir para um dos links daquela página, com chances iguais

Modelo do PageRank: Caminhada aleatória

- Imagine um navegador fazendo caminhada aleatória na web
  - Iniciar em uma página aleatória
  - A cada passo, sair da página atual e ir para um dos links daquela página, com chances iguais
- Depois de muita caminhada, cada página tem uma taxa de visita a longo prazo.

Modelo do PageRank: Caminhada aleatória

- Imagine um navegador fazendo caminhada aleatória na web
  - Iniciar em uma página aleatória
  - A cada passo, sair da página atual e ir para um dos links daquela página, com chances iguais
- Depois de muita caminhada, cada página tem uma taxa de visita a longo prazo.
- Essa taxa de visita de longo prazo é o PageRank da página.

Modelo do PageRank: Caminhada aleatória

- Imagine um navegador fazendo caminhada aleatória na web
  - Iniciar em uma página aleatória
  - A cada passo, sair da página atual e ir para um dos links daquela página, com chances iguais
- Depois de muita caminhada, cada página tem uma taxa de visita a longo prazo.
- Essa taxa de visita de longo prazo é o PageRank da página.
- PageRank = taxa de visita a longo prazo = probabilidade de estado estável

Formalização caminhada aleatória: Cadeia de Markov

Formalização caminhada aleatória: Cadeia de Markov

- A Cadeia de Markov consiste de N estados, mais uma matriz de probabilidade de transição P com $N \times N$ valores.

Formalização caminhada aleatória: Cadeia de Markov

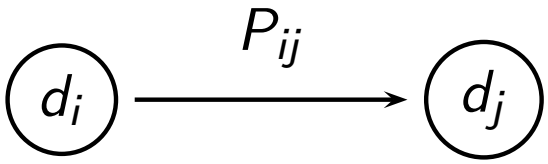
- A Cadeia de Markov consiste de N estados, mais uma matriz de probabilidade de transição P com $N \times N$ valores.
- estado = página

Formalização caminhada aleatória: Cadeia de Markov

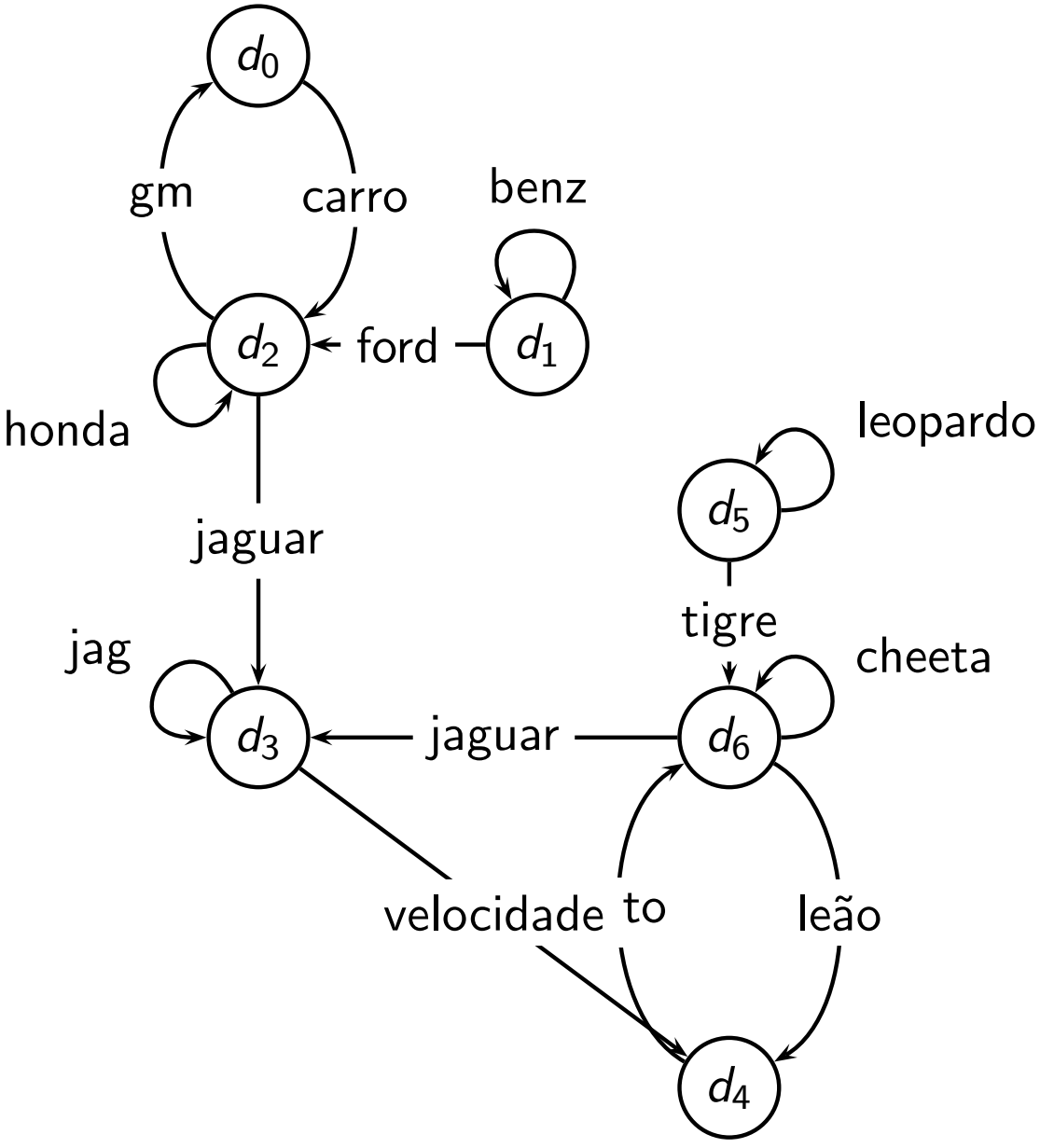
- A Cadeia de Markov consiste de N estados, mais uma matriz de probabilidade de transição P com $N \times N$ valores.
- estado = página
- A cada passo, estamos em uma das páginas

Formalização caminhada aleatória: Cadeia de Markov

- A Cadeia de Markov consiste de N estados, mais uma matriz de probabilidade de transição P com $N \times N$ valores.
- estado = página
- A cada passo, estamos em uma das páginas
- Para $1 \leq i, j \leq N$, o valor da matriz P_{ij} é a probabilidade de j ser a próxima página, dado que estamos atualmente na página i.
- Propriedade: $\sum_{j=1}^{N} P_{ij} = 1$



Exemplo grafo web



Exemplo: matriz de links

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0	0	1	0	0	0	0
d_1	0	1	1	0	0	0	0
d_2	1	0	1	1	0	0	0
d_3	0	0	0	1	1	0	0
d_4	0	0	0	0	0	0	1
d_5	0	0	0	0	0	1	1
d_6	0	0	0	1	1	0	1

Exemplo: matriz de probabilidades de transição P

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0.00	0.00	1.00	0.00	0.00	0.00	0.00
d_1	0.00	0.50	0.50	0.00	0.00	0.00	0.00
d_2	0.33	0.00	0.33	0.33	0.00	0.00	0.00
d_3	0.00	0.00	0.00	0.50	0.50	0.00	0.00
d_4	0.00	0.00	0.00	0.00	0.00	0.00	1.00
d_5	0.00	0.00	0.00	0.00	0.00	0.50	0.50
d_6	0.00	0.00	0.00	0.33	0.33	0.00	0.33

Taxa de visita a longo prazo

- PageRank = taxa de visita a longo prazo

Taxa de visita a longo prazo

- PageRank = taxa de visita a longo prazo
  - Taxa de visita a longo prazo da página d é a probabilidade que um navegador aleatório tem de estar na página d em um dado momento

Taxa de visita a longo prazo

- PageRank = taxa de visita a longo prazo
  - Taxa de visita a longo prazo da página d é a probabilidade que um navegador aleatório tem de estar na página d em um dado momento
- Quais são as propriedades que um grafo da web deve ter para a taxa de visita a longo prazo ser bem definida?

Taxa de visita a longo prazo

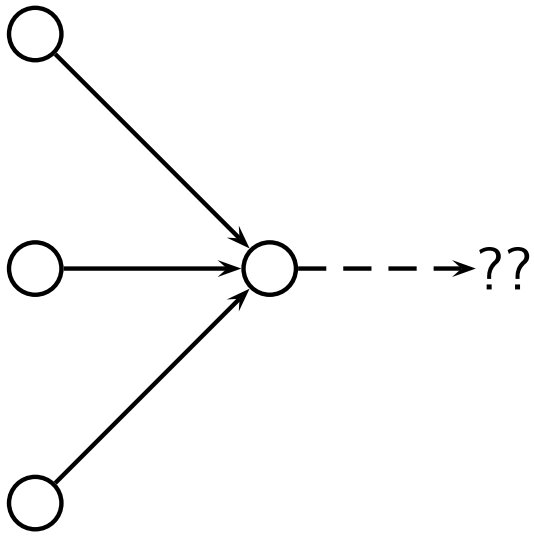
- PageRank = taxa de visita a longo prazo
  - Taxa de visita a longo prazo da página d é a probabilidade que um navegador aleatório tem de estar na página d em um dado momento
- Quais são as propriedades que um grafo da web deve ter para a taxa de visita a longo prazo ser bem definida?
- O grafo deve corresponder a uma cadeia de Markov ergódica.

Taxa de visita a longo prazo

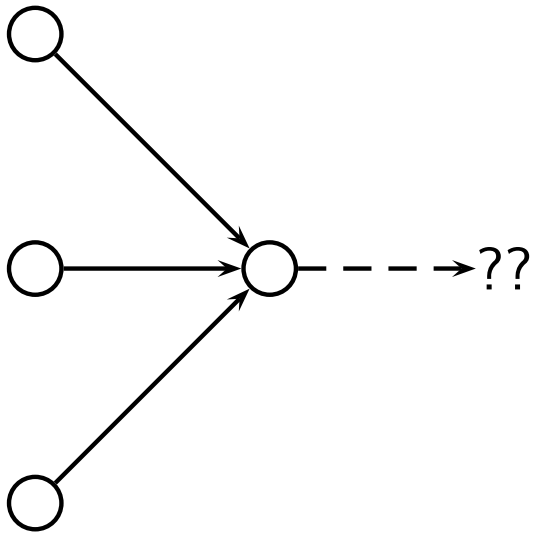
- PageRank = taxa de visita a longo prazo
  - Taxa de visita a longo prazo da página d é a probabilidade que um navegador aleatório tem de estar na página d em um dado momento
- Quais são as propriedades que um grafo da web deve ter para a taxa de visita a longo prazo ser bem definida?
- O grafo deve corresponder a uma cadeia de Markov ergódica.
- Caso especial: o grafo da web não deve ter becos sem saída

Becos sem saída

Becos sem saída

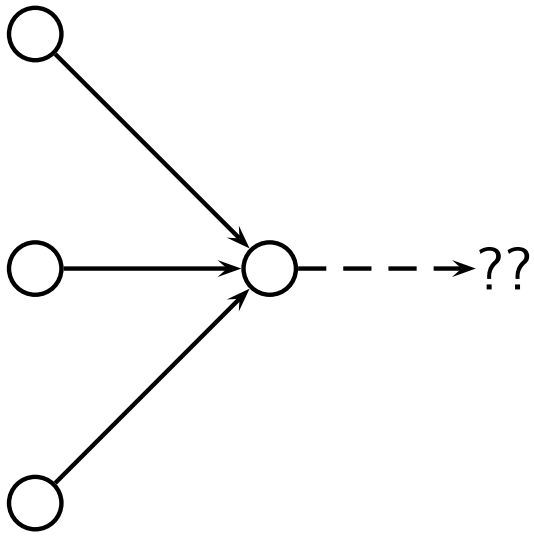


Becos sem saída



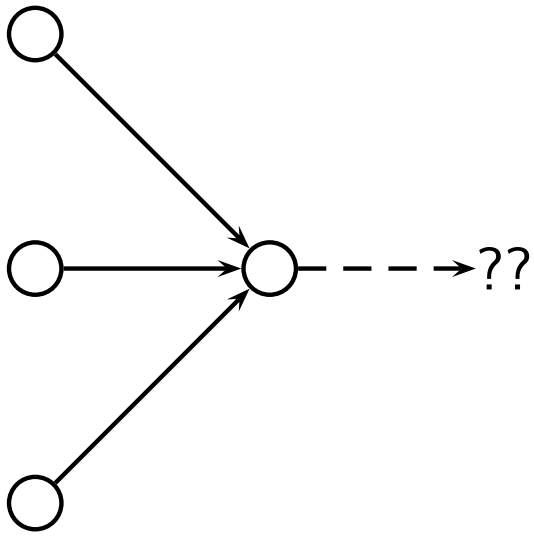
- A web é cheia de becos sem saída

Becos sem saída



- A web é cheia de becos sem saída
- Caminhada aleatória pode ficar travada em um beco sem saída

Becos sem saída



- A web é cheia de becos sem saída
- Caminhada aleatória pode ficar travada em um beco sem saída
- Se há becos sem saída, taxas de visita a longo prazo não são bem definidas

teletransporte – saída do caminho sem saída

- Em um beco sem saída, pular para uma página web aleatória com prob. $1/N$.

teletransporte – saída do caminho sem saída

- Em um beco sem saída, pular para uma página web aleatória com prob. $1/N$.
- Em um caminho normal, com probabilidade 10%, pular para uma página web aleatória (com probabilidade de $0.1/N$ para cada).

teletransporte – saída do caminho sem saída

- Em um beco sem saída, pular para uma página web aleatória com prob. $1/N$.
- Em um caminho normal, com probabilidade 10%, pular para uma página web aleatória (com probabilidade de $0.1/N$ para cada).
- Com probabilidade (90%), ir para um link aleatório

teletransporte – saída do caminho sem saída

- Em um beco sem saída, pular para uma página web aleatória com prob. $1/N$.
- Em um caminho normal, com probabilidade 10%, pular para uma página web aleatória (com probabilidade de $0.1/N$ para cada).
- Com probabilidade (90%), ir para um link aleatório
  - Exemplo: se uma página tem 4 links de saída: escolher um com probabilidade (1-0.10)/4=0.225

teletransporte – saída do caminho sem saída

- Em um beco sem saída, pular para uma página web aleatória com prob. $1/N$.
- Em um caminho normal, com probabilidade 10%, pular para uma página web aleatória (com probabilidade de $0.1/N$ para cada).
- Com probabilidade (90%), ir para um link aleatório
  - Exemplo: se uma página tem 4 links de saída: escolher um com probabilidade (1-0.10)/4=0.225
- 10% é um parâmetro, a taxa de teletransporte.

teletransporte – saída do caminho sem saída

- Em um beco sem saída, pular para uma página web aleatória com prob. $1/N$.
- Em um caminho normal, com probabilidade 10%, pular para uma página web aleatória (com probabilidade de $0.1/N$ para cada).
- Com probabilidade (90%), ir para um link aleatório
  - Exemplo: se uma página tem 4 links de saída: escolher um com probabilidade (1-0.10)/4=0.225
- 10% é um parâmetro, a taxa de teletransporte.
- Nota: "sair" de um beco sem saída é independente da taxa de teletransporte

Resultado de teletransporte

- Com teletransporte, não ficamos presos em um beco sem saída.

Resultado de teletransporte

- Com teletransporte, não ficamos presos em um beco sem saída.
- Mas mesmo sem becos sem saída, um grafo pode não ter taxa de visita a longo prazo bem definida.

Resultado de teletransporte

- Com teletransporte, não ficamos presos em um beco sem saída.
- Mas mesmo sem becos sem saída, um grafo pode não ter taxa de visita a longo prazo bem definida.
- É necessário que a cadeia de Markov seja ergódica.

Cadeias de Markov Ergódicas

- A cadeia de Markov é ergódica se e somente se for irredutível e aperiódica.

Cadeias de Markov Ergódicas

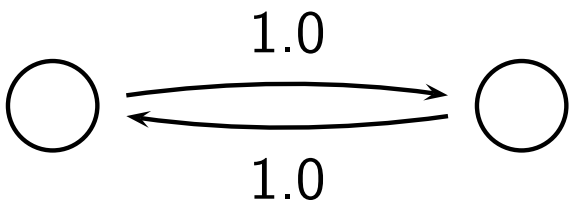
- A cadeia de Markov é ergódica se e somente se for irredutível e aperiódica.
- Irredutibilidade. Ideia: há um caminho entre quaisquer duas páginas

Cadeias de Markov Ergódicas

- A cadeia de Markov é ergódica se e somente se for irredutível e aperiódica.
- Irredutibilidade. Ideia: há um caminho entre quaisquer duas páginas
- Aperiodicidade. Ideia: Páginas não podem ser separadas de forma que o navegador aleatório seja obrigado a visitá-las em uma ordem específica.

Cadeias de Markov Ergódicas

- A cadeia de Markov é ergódica se e somente se for irredutível e aperiódica.
- Irredutibilidade. Ideia: há um caminho entre quaisquer duas páginas
- Aperiodicidade. Ideia: Páginas não podem ser separadas de forma que o navegador aleatório seja obrigado a visitá-las em uma ordem específica.
- Uma cadeia de Markov não-ergódica:



Cadeias de Markov Ergódicas

- Teorema: para uma cadeia de Markov ergódica, existe uma única taxa de visita a longo prazo para cada estado

Cadeias de Markov Ergódicas

- Teorema: para uma cadeia de Markov ergódica, existe uma única taxa de visita a longo prazo para cada estado
- Essa é a distribuição de probabilidades em estado estável.

Cadeias de Markov Ergódicas

- Teorema: para uma cadeia de Markov ergódica, existe uma única taxa de visita a longo prazo para cada estado
- Essa é a distribuição de probabilidades em estado estável.
- Em um longo período de tempo, visita-se cada estado na proporção dessa taxa

Cadeias de Markov Ergódicas

- Teorema: para uma cadeia de Markov ergódica, existe uma única taxa de visita a longo prazo para cada estado
- Essa é a distribuição de probabilidades em estado estável.
- Em um longo período de tempo, visita-se cada estado na proporção dessa taxa
- Não importa a página inicial

Cadeias de Markov Ergódicas

- Teorema: para uma cadeia de Markov ergódica, existe uma única taxa de visita a longo prazo para cada estado
- Essa é a distribuição de probabilidades em estado estável.
- Em um longo período de tempo, visita-se cada estado na proporção dessa taxa
- Não importa a página inicial
- teletransporte faz o grafo da web ser ergódico.

Cadeias de Markov Ergódicas

- Teorema: para uma cadeia de Markov ergódica, existe uma única taxa de visita a longo prazo para cada estado
- Essa é a distribuição de probabilidades em estado estável.
- Em um longo período de tempo, visita-se cada estado na proporção dessa taxa
- Não importa a página inicial
- teletransporte faz o grafo da web ser ergódico.
- $\Rightarrow$ Grafo-Web+teletransporte tem uma distribuição de probabilidades de estado estável.

Cadeias de Markov Ergódicas

- Teorema: para uma cadeia de Markov ergódica, existe uma única taxa de visita a longo prazo para cada estado
- Essa é a distribuição de probabilidades em estado estável.
- Em um longo período de tempo, visita-se cada estado na proporção dessa taxa
- Não importa a página inicial
- teletransporte faz o grafo da web ser ergódico.
- $\Rightarrow$ Grafo-Web+teletransporte tem uma distribuição de probabilidades de estado estável.
- $\Rightarrow$ Cada página no grafo-web+teletransporte tem um valor de PageRank.

Onde estamos

- Sabemos como ter certeza que temos um PageRank para cada página

Onde estamos

- Sabemos como ter certeza que temos um PageRank para cada página
- Como calcular?

Formalização de “visitas”: vetor de probabilidades

- Um vetor de probabilidades (vetor-linha) $\vec{x} = (x_1, \ldots, x_N)$ diz onde o navegador aleatório está em qualquer momento

Formalização de "visitas": vetor de probabilidades

- Um vetor de probabilidades (vetor-linha) $\vec{x} = (x_1, \ldots, x_N)$ diz onde o navegador aleatório está em qualquer momento
- Exemplo:

(	0	0	0	...	1	...	0	0	0	)
	1	2	3	...	i	...	N-2	N-1	N	

Formalização de "visitas": vetor de probabilidades

- Um vetor de probabilidades (vetor-linha) $\vec{x} = (x_1, \ldots, x_N)$ diz onde o navegador aleatório está em qualquer momento
- Exemplo:

(	0	0	0	...	1	...	0	0	0	)
	1	2	3	...	i	...	N-2	N-1	N	

- Mais geral: caminhada aleatória está na página i com probabilidade x_i.

Formalização de "visitas": vetor de probabilidades

- Um vetor de probabilidades (vetor-linha) $\vec{x} = (x_1, \ldots, x_N)$ diz onde o navegador aleatório está em qualquer momento
- Exemplo:

(	0	0	0	...	1	...	0	0	0	)
	1	2	3	...	i	...	N-2	N-1	N	

- Mais geral: caminhada aleatória está na página i com probabilidade x_i.
- Exemplo:

(	0.05	0.01	0.0	...	0.2	...	0.01	0.05	0.03	)
	1	2	3	...	i	...	N-2	N-1	N	

Formalização de "visitas": vetor de probabilidades

- Um vetor de probabilidades (vetor-linha) $\vec{x} = (x_1, \ldots, x_N)$ diz onde o navegador aleatório está em qualquer momento
- Exemplo:

(	0	0	0	...	1	...	0	0	0	)
	1	2	3	...	i	...	N-2	N-1	N	

- Mais geral: caminhada aleatória está na página i com probabilidade x_i.
- Exemplo:

(	0.05	0.01	0.0	...	0.2	...	0.01	0.05	0.03	)
	1	2	3	...	i	...	N-2	N-1	N	

- $\sum x_i = 1$

Mudança no vetor de probabilidades

- Se o vetor de probabilidades é $\vec{x} = (x_1, \ldots, x_N)$ neste passo, como será no próximo?

Mudança no vetor de probabilidades

- Se o vetor de probabilidades é $\vec{x} = (x_1, \ldots, x_N)$ neste passo, como será no próximo?
- Lembre-se que a linha i da matriz de transição de probabilidades P nos diz para onde iremos a partir do estado i.

Mudança no vetor de probabilidades

- Se o vetor de probabilidades é $\vec{x} = (x_1, \ldots, x_N)$ neste passo, como será no próximo?
- Lembre-se que a linha i da matriz de transição de probabilidades P nos diz para onde iremos a partir do estado i.
- Então a partir de $\vec{x}$, o próximo estado é distribuído com as probabilidades em $\vec{x}P$.

Notação de estado estável

- O estado estável na notação de vetor é simplesmente um vetor $\vec{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ de probabilidades

Notação de estado estável

- O estado estável na notação de vetor é simplesmente um vetor $\vec{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ de probabilidades
- Usamos $\vec{\pi}$ para diferenciar da notação de vetor de probabilidades $\vec{x}$.

Notação de estado estável

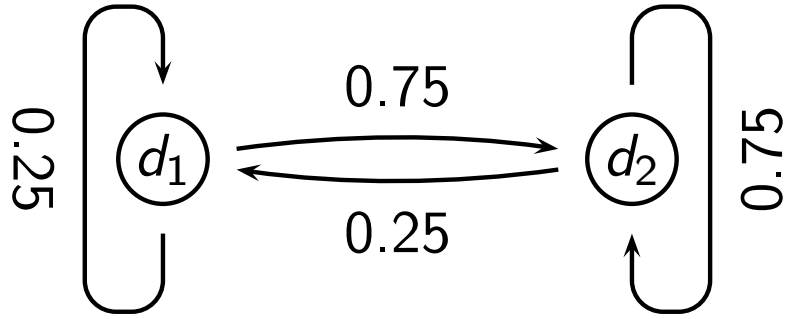
- O estado estável na notação de vetor é simplesmente um vetor $\vec{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ de probabilidades
- Usamos $\vec{\pi}$ para diferenciar da notação de vetor de probabilidades $\vec{x}$.
- π_i é a taxa de visita a longo prazo (ou PageRank) da página i.

Notação de estado estável

- O estado estável na notação de vetor é simplesmente um vetor $\vec{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ de probabilidades
- Usamos $\vec{\pi}$ para diferenciar da notação de vetor de probabilidades $\vec{x}$.
- π_i é a taxa de visita a longo prazo (ou PageRank) da página i.
- Podemos considerar o PageRank como um grande vetor – um valor por página

Exemplo: distribuição de estado-estável

- O que é o PageRank / estado-estável nesse exemplo?



Exemplo: estado estável

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$	
			$P_{11} = 0.25$ $P_{12} = 0.75$
			$P_{21} = 0.25$ $P_{22} = 0.75$
t_0			
t_1			

Vetor PageRank $= \vec{\pi} = (\pi_1, \pi_2) = (0.25, 0.75)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Exemplo: estado estável

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$		
			$P_{11} = 0.25$	$P_{12} = 0.75$
			$P_{21} = 0.25$	$P_{22} = 0.75$
t_0	0.25	0.75		
t_1				

Vetor PageRank $= \vec{\pi} = (\pi_1, \pi_2) = (0.25, 0.75)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Exemplo: estado estável

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$		
			$P_{11} = 0.25$ $P_{21} = 0.25$	$P_{12} = 0.75$ $P_{22} = 0.75$
t_0	0.25	0.75	0.25	0.75
t_1				

Vetor PageRank $= \vec{\pi} = (\pi_1, \pi_2) = (0.25, 0.75)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Exemplo: estado estável

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$		
			$P_{11} = 0.25$	$P_{12} = 0.75$
			$P_{21} = 0.25$	$P_{22} = 0.75$
t_0	0.25	0.75	0.25	0.75
t_1	0.25	0.75		

Vetor PageRank $= \vec{\pi} = (\pi_1, \pi_2) = (0.25, 0.75)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Exemplo: estado estável

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$		
			$P_{11} = 0.25$	$P_{12} = 0.75$
			$P_{21} = 0.25$	$P_{22} = 0.75$
t_0	0.25	0.75	0.25	0.75
t_1	0.25	0.75	(convergência)	

Vetor PageRank $= \vec{\pi} = (\pi_1, \pi_2) = (0.25, 0.75)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Como obter o vetor de estado estável?

- Como calcular PageRank?

Como obter o vetor de estado estável?

- Como calcular PageRank?
- Vetor PageRank: $\vec{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$, ...

Como obter o vetor de estado estável?

- Como calcular PageRank?
- Vetor PageRank: $\vec{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$, ...
- ...se a distribuição neste passo é $\vec{x}$, então a distribuição no próximo passo é $\vec{x}P$.

Como obter o vetor de estado estável?

- Como calcular PageRank?
- Vetor PageRank: $\vec{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$, . . .
- . . . se a distribuição neste passo é $\vec{x}$, então a distribuição no próximo passo é $\vec{x}P$.
- Mas $\vec{\pi}$ é o estado estável!

Como obter o vetor de estado estável?

- Como calcular PageRank?
- Vetor PageRank: $\vec{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$, . . .
- . . . se a distribuição neste passo é $\vec{x}$, então a distribuição no próximo passo é $\vec{x}P$.
- Mas $\vec{\pi}$ é o estado estável!
- Então: $\vec{\pi} = \vec{\pi}P$

Como obter o vetor de estado estável?

- Como calcular PageRank?
- Vetor PageRank: $\vec{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$, . . .
- . . . se a distribuição neste passo é $\vec{x}$, então a distribuição no próximo passo é $\vec{x}P$.
- Mas $\vec{\pi}$ é o estado estável!
- Então: $\vec{\pi} = \vec{\pi}P$
- Resolvendo essa equação nos dá $\vec{\pi}$.

Como obter o vetor de estado estável?

- Como calcular PageRank?
- Vetor PageRank: $\vec{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$, ...
- ...se a distribuição neste passo é $\vec{x}$, então a distribuição no próximo passo é $\vec{x}P$.
- Mas $\vec{\pi}$ é o estado estável!
- Então: $\vec{\pi} = \vec{\pi}P$
- Resolvendo essa equação nos dá $\vec{\pi}$.
- $\vec{\pi}$ é o auto-vetor principal esquerdo para P ...

Como obter o vetor de estado estável?

- Como calcular PageRank?
- Vetor PageRank: $\vec{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$, ...
- ...se a distribuição neste passo é $\vec{x}$, então a distribuição no próximo passo é $\vec{x}P$.
- Mas $\vec{\pi}$ é o estado estável!
- Então: $\vec{\pi} = \vec{\pi}P$
- Resolvendo essa equação nos dá $\vec{\pi}$.
- $\vec{\pi}$ é o auto-vetor principal esquerdo para P ...
- ...isto é, $\vec{\pi}$ é o auto-vetor esquerdo com maior auto-valor

Como obter o vetor de estado estável?

- Como calcular PageRank?
- Vetor PageRank: $\vec{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$, ...
- ... se a distribuição neste passo é $\vec{x}$, então a distribuição no próximo passo é $\vec{x}P$.
- Mas $\vec{\pi}$ é o estado estável!
- Então: $\vec{\pi} = \vec{\pi}P$
- Resolvendo essa equação nos dá $\vec{\pi}$.
- $\vec{\pi}$ é o auto-vetor principal esquerdo para P ...
- ... isto é, $\vec{\pi}$ é o auto-vetor esquerdo com maior auto-valor
- Todas as matrizes de probabilidades de transição tem o maior auto-valor igual a 1.

Um modo de calcular o PageRank $\vec{\pi}$

- Iniciar com qualquer distribuição $\vec{x}$, e.g., distribuição uniforme

Um modo de calcular o PageRank $\vec{\pi}$

- Iniciar com qualquer distribuição $\vec{x}$, e.g., distribuição uniforme
- Depois de um passo, estamos em $\vec{x}P$.

Um modo de calcular o PageRank $\vec{\pi}$

- Iniciar com qualquer distribuição $\vec{x}$, e.g., distribuição uniforme
- Depois de um passo, estamos em $\vec{x}P$.
- Depois de dois passos, estamos em $\vec{x}P^2$.

Um modo de calcular o PageRank $\vec{\pi}$

- Iniciar com qualquer distribuição $\vec{x}$, e.g., distribuição uniforme
- Depois de um passo, estamos em $\vec{x}P$.
- Depois de dois passos, estamos em $\vec{x}P^2$.
- Depois de k passos, estamos em $\vec{x}P^k$.

Um modo de calcular o PageRank $\vec{\pi}$

- Iniciar com qualquer distribuição $\vec{x}$, e.g., distribuição uniforme
- Depois de um passo, estamos em $\vec{x}P$.
- Depois de dois passos, estamos em $\vec{x}P^2$.
- Depois de k passos, estamos em $\vec{x}P^k$.
- Algoritmo: multiplicar $\vec{x}$ por potências de P até convergência

Um modo de calcular o PageRank $\vec{\pi}$

- Iniciar com qualquer distribuição $\vec{x}$, e.g., distribuição uniforme
- Depois de um passo, estamos em $\vec{x}P$.
- Depois de dois passos, estamos em $\vec{x}P^2$.
- Depois de k passos, estamos em $\vec{x}P^k$.
- Algoritmo: multiplicar $\vec{x}$ por potências de P até convergência
- Método da potência

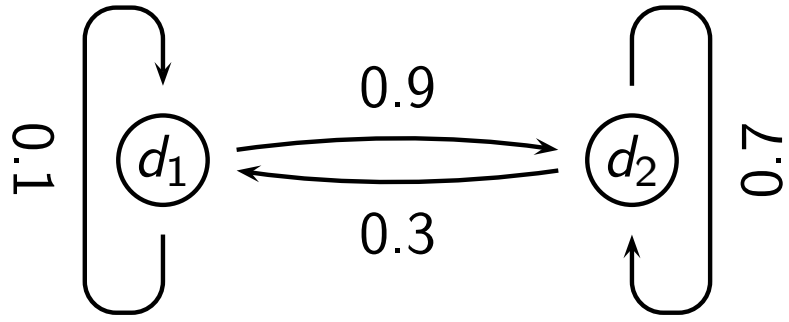
Um modo de calcular o PageRank $\vec{\pi}$

- Iniciar com qualquer distribuição $\vec{x}$, e.g., distribuição uniforme
- Depois de um passo, estamos em $\vec{x}P$.
- Depois de dois passos, estamos em $\vec{x}P^2$.
- Depois de k passos, estamos em $\vec{x}P^k$.
- Algoritmo: multiplicar $\vec{x}$ por potências de P até convergência
- Método da potência
- Eventualmente chegamos no estado estável $\vec{\pi}$.

Exemplo: método da potência

Exemplo: método da potência

- O que é o PageRank / estado estável neste exemplo?



Calculando PageRank: método das potências

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$	$P_{12} = 0.9$	
			$P_{21} = 0.3$	$P_{22} = 0.7$	
t_0					$= \vec{x}P$
t_1					$= \vec{x}P^2$
t_2					$= \vec{x}P^3$
t_3					$= \vec{x}P^4$
					$\ldots$
t_∞					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Calculando PageRank: método das potências

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$	$P_{12} = 0.9$	
			$P_{21} = 0.3$	$P_{22} = 0.7$	
t_0	0	1			$= \vec{x}P$
t_1					$= \vec{x}P^2$
t_2					$= \vec{x}P^3$
t_3					$= \vec{x}P^4$
					$\ldots$
t_∞					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Calculando PageRank: método das potências

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1					$= \vec{x}P^2$
t_2					$= \vec{x}P^3$
t_3					$= \vec{x}P^4$
					$\ldots$
t_∞					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Calculando PageRank: método das potências

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7			$= \vec{x}P^2$
t_2					$= \vec{x}P^3$
t_3					$= \vec{x}P^4$
					$\ldots$
t_∞					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Calculando PageRank: método das potências

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
t_2					$= \vec{x}P^3$
t_3					$= \vec{x}P^4$
					$\ldots$
t_∞					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Calculando PageRank: método das potências

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
t_2	0.24	0.76			$= \vec{x}P^3$
t_3					$= \vec{x}P^4$
					$\ldots$
t_∞					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Calculando PageRank: método das potências

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
t_2	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
t_3					$= \vec{x}P^4$
					$\ldots$
t_∞					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Calculando PageRank: método das potências

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
t_2	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
t_3	0.252	0.748			$= \vec{x}P^4$
					$\ldots$
t_∞					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Calculando PageRank: método das potências

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
t_2	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
t_3	0.252	0.748	0.2496	0.7504	$= \vec{x}P^4$
					$\ldots$
t_∞					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Calculando PageRank: método das potências

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
t_2	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
t_3	0.252	0.748	0.2496	0.7504	$= \vec{x}P^4$
			...		...
t_∞					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Calculando PageRank: método das potências

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
t_2	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
t_3	0.252	0.748	0.2496	0.7504	$= \vec{x}P^4$
			...		...
t_∞	0.25	0.75			$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Calculando PageRank: método das potências

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
t_2	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
t_3	0.252	0.748	0.2496	0.7504	$= \vec{x}P^4$
			...		...
t_∞	0.25	0.75	0.25	0.75	$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Calculando PageRank: método das potências

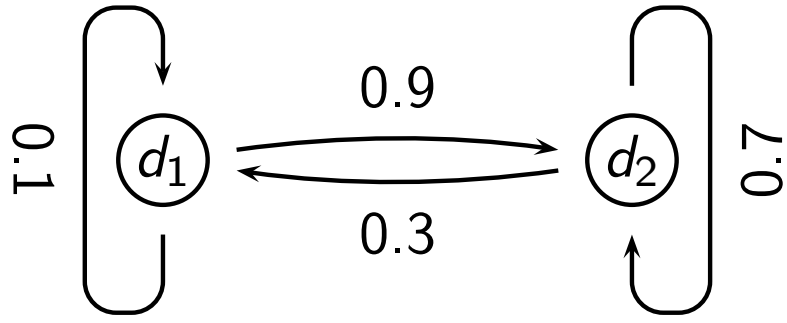
	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
t_2	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
t_3	0.252	0.748	0.2496	0.7504	$= \vec{x}P^4$
			...		...
t_∞	0.25	0.75	0.25	0.75	$= \vec{x}P^\infty$

Vetor PageRank $= \vec{\pi} = (\pi_1, \pi_2) = (0.25, 0.75)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

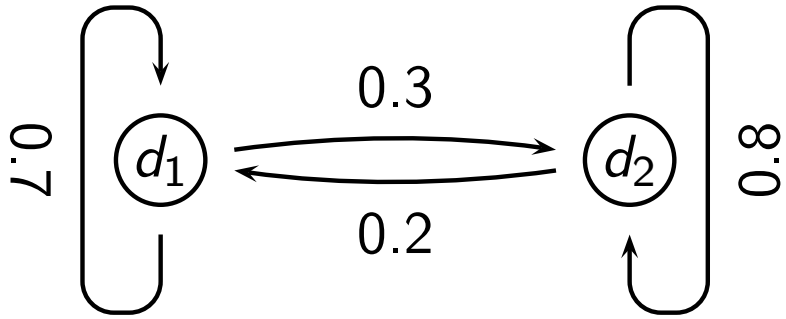
Exemplo: método da potência

- O que é o PageRank / estado estável neste exemplo?



- A distribuição de estado estável (= os PageRanks) nesse exemplo são 0.25 para d_1 e 0.75 para d_2.

Exercício: calcular o PageRank usando o método das potências



Solução

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$	
			$P_{11} = 0.7$ $P_{12} = 0.3$
			$P_{21} = 0.2$ $P_{22} = 0.8$
t_0			
t_1			
t_2			
t_3			
t_∞			

Vetor PageRank $= \vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Solução

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$		
			$P_{11} = 0.7$	$P_{12} = 0.3$
			$P_{21} = 0.2$	$P_{22} = 0.8$
t_0	0	1		
t_1				
t_2				
t_3				
t_∞				

Vetor PageRank $= \vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Solução

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$		
			$P_{11} = 0.7$	$P_{12} = 0.3$
			$P_{21} = 0.2$	$P_{22} = 0.8$
t_0	0	1	0.2	0.8
t_1				
t_2				
t_3				
t_∞				

Vetor PageRank $= \vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Solução

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$		
			$P_{11} = 0.7$	$P_{12} = 0.3$
			$P_{21} = 0.2$	$P_{22} = 0.8$
t_0	0	1	0.2	0.8
t_1	0.2	0.8		
t_2				
t_3				
t_∞				

Vetor PageRank $= \vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Solução

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$		
			$P_{11} = 0.7$ $P_{21} = 0.2$	$P_{12} = 0.3$ $P_{22} = 0.8$
t_0	0	1	0.2	0.8
t_1	0.2	0.8	0.3	0.7
t_2				
t_3				
t_∞				

Vetor PageRank $= \vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Solução

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$		
			$P_{11} = 0.7$ $P_{21} = 0.2$	$P_{12} = 0.3$ $P_{22} = 0.8$
t_0	0	1	0.2	0.8
t_1	0.2	0.8	0.3	0.7
t_2	0.3	0.7		
t_3				
t_∞				

Vetor PageRank $= \vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Solução

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$		
			$P_{11} = 0.7$ $P_{21} = 0.2$	$P_{12} = 0.3$ $P_{22} = 0.8$
t_0	0	1	0.2	0.8
t_1	0.2	0.8	0.3	0.7
t_2	0.3	0.7	0.35	0.65
t_3				
t_∞				

Vetor PageRank $= \vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Solução

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$		
			$P_{11} = 0.7$ $P_{21} = 0.2$	$P_{12} = 0.3$ $P_{22} = 0.8$
t_0	0	1	0.2	0.8
t_1	0.2	0.8	0.3	0.7
t_2	0.3	0.7	0.35	0.65
t_3	0.35	0.65		
t_∞				

Vetor PageRank $= \vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Solução

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$		
			$P_{11} = 0.7$ $P_{21} = 0.2$	$P_{12} = 0.3$ $P_{22} = 0.8$
t_0	0	1	0.2	0.8
t_1	0.2	0.8	0.3	0.7
t_2	0.3	0.7	0.35	0.65
t_3	0.35	0.65	0.375	0.625
t_∞				

Vetor PageRank $= \vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Solução

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$		
			$P_{11} = 0.7$ $P_{21} = 0.2$	$P_{12} = 0.3$ $P_{22} = 0.8$
t_0	0	1	0.2	0.8
t_1	0.2	0.8	0.3	0.7
t_2	0.3	0.7	0.35	0.65
t_3	0.35	0.65	0.375	0.625
				...
t_∞				

Vetor PageRank $= \vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Solução

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$		
			$P_{11} = 0.7$ $P_{21} = 0.2$	$P_{12} = 0.3$ $P_{22} = 0.8$
t_0	0	1	0.2	0.8
t_1	0.2	0.8	0.3	0.7
t_2	0.3	0.7	0.35	0.65
t_3	0.35	0.65	0.375	0.625
				...
t_∞	0.4	0.6		

Vetor PageRank $= \vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Solução

	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$		
			$P_{11} = 0.7$ $P_{21} = 0.2$	$P_{12} = 0.3$ $P_{22} = 0.8$
t_0	0	1	0.2	0.8
t_1	0.2	0.8	0.3	0.7
t_2	0.3	0.7	0.35	0.65
t_3	0.35	0.65	0.375	0.625
				...
t_∞	0.4	0.6	0.4	0.6

Vetor PageRank $= \vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

PageRank: resumo

- Preprocessamento

PageRank: resumo

- Preprocessamento
  - Dado um grafo de links, construir matriz P

PageRank: resumo

- Preprocessamento
  - Dado um grafo de links, construir matriz P
  - Aplicar teletransporte

PageRank: resumo

- Preprocessamento
  - Dado um grafo de links, construir matriz P
  - Aplicar teletransporte
  - A partir da matriz modificada, calcular $\vec{\pi}$

PageRank: resumo

- Preprocessamento
  - Dado um grafo de links, construir matriz P
  - Aplicar teletransporte
  - A partir da matriz modificada, calcular $\vec{\pi}$
  - $\vec{\pi}_i$ é o PageRank da página i.

PageRank: resumo

- Preprocessamento
  - Dado um grafo de links, construir matriz P
  - Aplicar teletransporte
  - A partir da matriz modificada, calcular $\vec{\pi}$
  - $\vec{\pi}_i$ é o PageRank da página i.
- Processamento da consulta

PageRank: resumo

- Preprocessamento
  - Dado um grafo de links, construir matriz P
  - Aplicar teletransporte
  - A partir da matriz modificada, calcular $\vec{\pi}$
  - $\vec{\pi}_i$ é o PageRank da página i.
- Processamento da consulta
  - Recuperar páginas satisfazendo a consulta

PageRank: resumo

- Preprocessamento
  - Dado um grafo de links, construir matriz P
  - Aplicar teletransporte
  - A partir da matriz modificada, calcular $\vec{\pi}$
  - $\vec{\pi}_i$ é o PageRank da página i.
- Processamento da consulta
  - Recuperar páginas satisfazendo a consulta
  - Rankear usando o PageRank

PageRank: resumo

- Preprocessamento
  - Dado um grafo de links, construir matriz P
  - Aplicar teletransporte
  - A partir da matriz modificada, calcular $\vec{\pi}$
  - $\vec{\pi}_i$ é o PageRank da página i.
- Processamento da consulta
  - Recuperar páginas satisfazendo a consulta
  - Rankear usando o PageRank
  - Retornar lista rerankeada para o usuário

Problemas com PageRank

- Usuários reais não são aleatórios

Problemas com PageRank

- Usuários reais não são aleatórios
  - Exemplos de navegação não-aleatória: botão voltar, favoritos, diretórios – e buscadores

Problemas com PageRank

- Usuários reais não são aleatórios
  - Exemplos de navegação não-aleatória: botão voltar, favoritos, diretórios – e buscadores
  - $\rightarrow$ Modelo de Markov não é um bom modelo de navegação na web

Problemas com PageRank

- Usuários reais não são aleatórios
  - Exemplos de navegação não-aleatória: botão voltar, favoritos, diretórios – e buscadores
  - $\rightarrow$ Modelo de Markov não é um bom modelo de navegação na web
  - Mas é bom suficiente para nossos interesses

Problemas com PageRank

- Usuários reais não são aleatórios
  - Exemplos de navegação não-aleatória: botão voltar, favoritos, diretórios – e buscadores
  - $\rightarrow$ Modelo de Markov não é um bom modelo de navegação na web
  - Mas é bom suficiente para nossos interesses
- Ranking PageRank (como descrito anteriormente) pode produzir resultados ruins para muitas páginas

Problemas com PageRank

- Usuários reais não são aleatórios
  - Exemplos de navegação não-aleatória: botão voltar, favoritos, diretórios – e buscadores
  - $\rightarrow$ Modelo de Markov não é um bom modelo de navegação na web
  - Mas é bom suficiente para nossos interesses
- Ranking PageRank (como descrito anteriormente) pode produzir resultados ruins para muitas páginas
  - Considere a consulta [serviço vídeo]

Problemas com PageRank

- Usuários reais não são aleatórios
  - Exemplos de navegação não-aleatória: botão voltar, favoritos, diretórios – e buscadores
  - $\rightarrow$ Modelo de Markov não é um bom modelo de navegação na web
  - Mas é bom suficiente para nossos interesses
- Ranking PageRank (como descrito anteriormente) pode produzir resultados ruins para muitas páginas
  - Considere a consulta [serviço vídeo]
  - A home page Yahoo (i) tem PageRank alto e (ii) contém ambos *vídeo* e *serviço*.

Problemas com PageRank

- Usuários reais não são aleatórios
  - Exemplos de navegação não-aleatória: botão voltar, favoritos, diretórios – e buscadores
  - $\rightarrow$ Modelo de Markov não é um bom modelo de navegação na web
  - Mas é bom suficiente para nossos interesses
- Ranking PageRank (como descrito anteriormente) pode produzir resultados ruins para muitas páginas
  - Considere a consulta [serviço vídeo]
  - A home page Yahoo (i) tem PageRank alto e (ii) contém ambos *vídeo* e *serviço*.
  - Se rankeamos os resultados Booleanos de acordo com o PageRank, então o Yahoo seria o melhor ranking

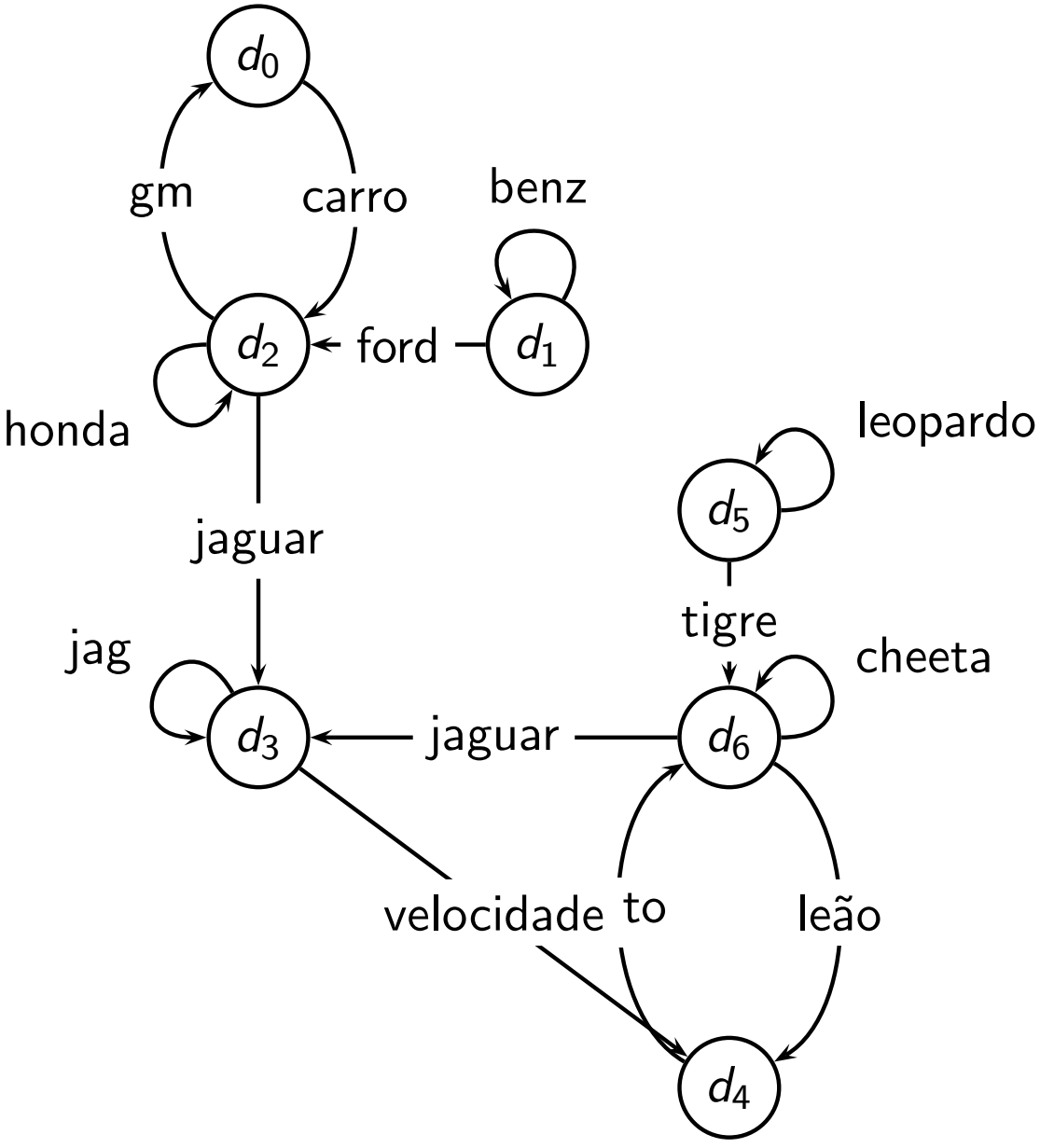
Problemas com PageRank

- Usuários reais não são aleatórios
  - Exemplos de navegação não-aleatória: botão voltar, favoritos, diretórios – e buscadores
  - $\rightarrow$ Modelo de Markov não é um bom modelo de navegação na web
  - Mas é bom suficiente para nossos interesses
- Ranking PageRank (como descrito anteriormente) pode produzir resultados ruins para muitas páginas
  - Considere a consulta [serviço vídeo]
  - A home page Yahoo (i) tem PageRank alto e (ii) contém ambos *vídeo* e *serviço*.
  - Se rankeamos os resultados Booleanos de acordo com o PageRank, então o Yahoo seria o melhor ranking
  - Não desejável

Problemas com PageRank

- Usuários reais não são aleatórios
  - Exemplos de navegação não-aleatória: botão voltar, favoritos, diretórios – e buscadores
  - $\rightarrow$ Modelo de Markov não é um bom modelo de navegação na web
  - Mas é bom suficiente para nossos interesses
- Ranking PageRank (como descrito anteriormente) pode produzir resultados ruins para muitas páginas
  - Considere a consulta [serviço vídeo]
  - A home page Yahoo (i) tem PageRank alto e (ii) contém ambos *vídeo* e *serviço*.
  - Se rankeamos os resultados Booleanos de acordo com o PageRank, então o Yahoo seria o melhor ranking
  - Não desejável
- Na prática: rankear de acordo com a combinação ponderada da comparação do texto do documento, do texto âncora, do PageRank e outros fatores

Exemplo grafo web



Matriz de probabilidades de transição

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0.00	0.00	1.00	0.00	0.00	0.00	0.00
d_1	0.00	0.50	0.50	0.00	0.00	0.00	0.00
d_2	0.33	0.00	0.33	0.33	0.00	0.00	0.00
d_3	0.00	0.00	0.00	0.50	0.50	0.00	0.00
d_4	0.00	0.00	0.00	0.00	0.00	0.00	1.00
d_5	0.00	0.00	0.00	0.00	0.00	0.50	0.50
d_6	0.00	0.00	0.00	0.33	0.33	0.00	0.33

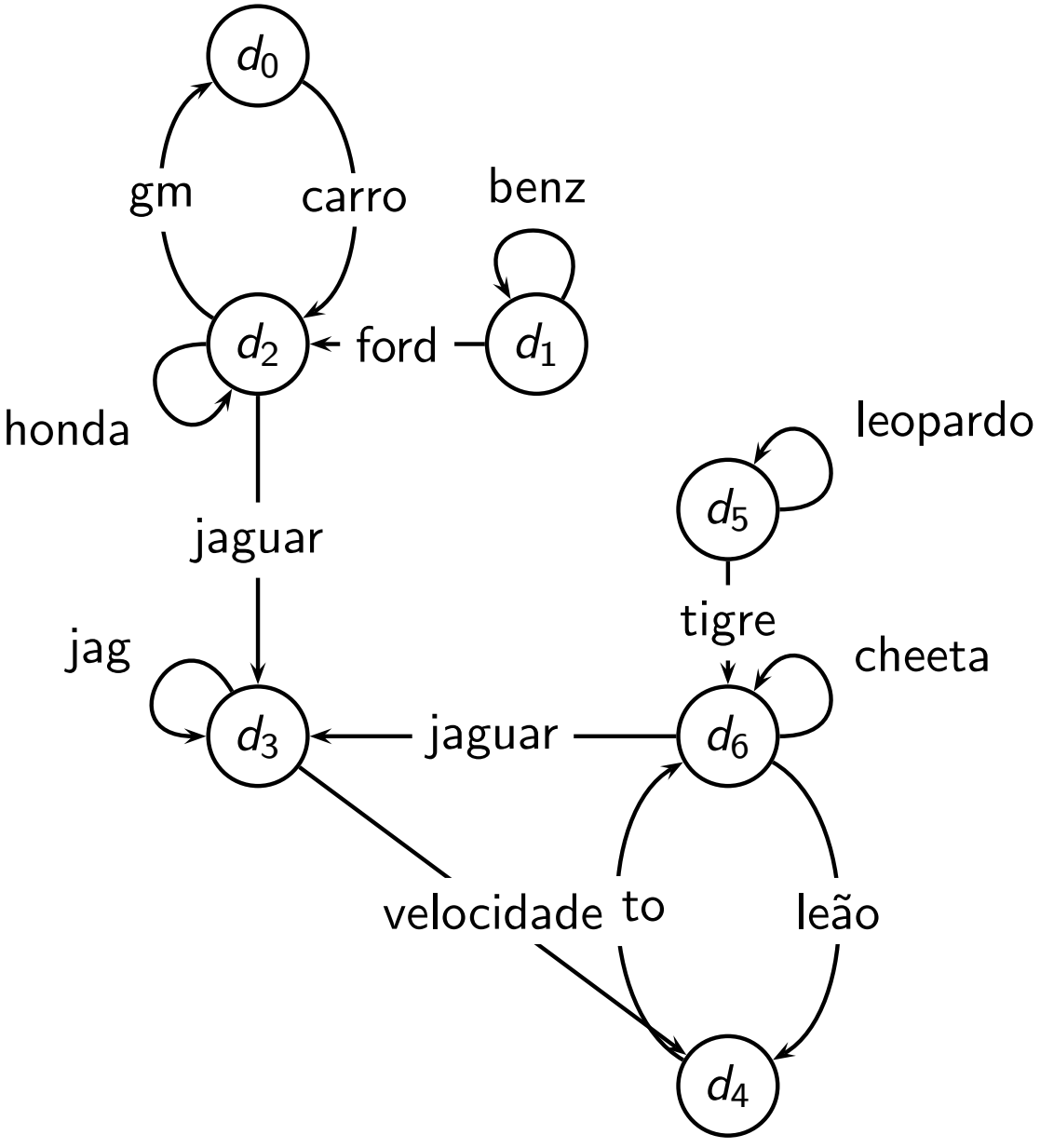
Matriz de transição com teletransporte

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0.02	0.02	0.88	0.02	0.02	0.02	0.02
d_1	0.02	0.45	0.45	0.02	0.02	0.02	0.02
d_2	0.31	0.02	0.31	0.31	0.02	0.02	0.02
d_3	0.02	0.02	0.02	0.45	0.45	0.02	0.02
d_4	0.02	0.02	0.02	0.02	0.02	0.02	0.88
d_5	0.02	0.02	0.02	0.02	0.02	0.45	0.45
d_6	0.02	0.02	0.02	0.31	0.31	0.02	0.31

Vetores do método de potências $\vec{x}P^k$

	$\vec{x}$	$\vec{x}P^1$	$\vec{x}P^2$	$\vec{x}P^3$	$\vec{x}P^4$	$\vec{x}P^5$	$\vec{x}P^6$	$\vec{x}P^7$	$\vec{x}P^8$	$\vec{x}P^9$	$\vec{x}P^{10}$	$\vec{x}P^{11}$	$\vec{x}P^{12}$	$\vec{x}P^{13}$
d_0	0.14	0.06	0.09	0.07	0.07	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05
d_1	0.14	0.08	0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
d_2	0.14	0.25	0.18	0.17	0.15	0.14	0.13	0.12	0.12	0.12	0.12	0.11	0.11	0.11
d_3	0.14	0.16	0.23	0.24	0.24	0.24	0.24	0.25	0.25	0.25	0.25	0.25	0.25	0.25
d_4	0.14	0.12	0.16	0.19	0.19	0.20	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21
d_5	0.14	0.08	0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
d_6	0.14	0.25	0.23	0.25	0.27	0.28	0.29	0.29	0.30	0.30	0.30	0.30	0.31	0.31

Exemplo grafo web



	PageRank
d_0	0.05
d_1	0.04
d_2	0.11
d_3	0.25
d_4	0.21
d_5	0.04
d_6	0.31

PageRank(d2)< PageRank(d6): porquê?

Importância do PageRank

- Alegação frequente: PageRank é o componente mais importante do ranking de páginas na web

Importância do PageRank

- Alegação frequente: PageRank é o componente mais importante do ranking de páginas na web
- A realidade:

Importância do PageRank

- Alegação frequente: PageRank é o componente mais importante do ranking de páginas na web
- A realidade:
  - Há vários componentes que são pelo menos tão importantes: e.g., texto âncora, expressões, proximidade, índices em camadas ...

Importância do PageRank

- Alegação frequente: PageRank é o componente mais importante do ranking de páginas na web
- A realidade:
  - Há vários componentes que são pelo menos tão importantes: e.g., texto âncora, expressões, proximidade, índices em camadas ...
  - Diz-se que o PageRank no formato original (como mostrado aqui) hoje tem um impacto baixo no ranking

Importância do PageRank

- Alegação frequente: PageRank é o componente mais importante do ranking de páginas na web
- A realidade:
  - Há vários componentes que são pelo menos tão importantes: e.g., texto âncora, expressões, proximidade, índices em camadas ...
  - Diz-se que o PageRank no formato original (como mostrado aqui) hoje tem um impacto baixo no ranking
  - Porém, variantes do PageRank são ainda essenciais para a recuperação de páginas web

Importância do PageRank

- Alegação frequente: PageRank é o componente mais importante do ranking de páginas na web
- A realidade:
  - Há vários componentes que são pelo menos tão importantes: e.g., texto âncora, expressões, proximidade, índices em camadas ...
  - Diz-se que o PageRank no formato original (como mostrado aqui) hoje tem um impacto baixo no ranking
  - Porém, variantes do PageRank são ainda essenciais para a recuperação de páginas web
  - Lutar contra spam baseado em links é difícil e crucial

Conteúdo

HITS – Hyperlink-Induced Topic Search

- Premissa: dois tipos de relevância na web

HITS – Hyperlink-Induced Topic Search

- Premissa: dois tipos de relevância na web
- Tipo 1: Hubs. Uma página hub é uma boa lista de [links para páginas adequadas à necessidade de informação].

HITS – Hyperlink-Induced Topic Search

- Premissa: dois tipos de relevância na web
- Tipo 1: Hubs. Uma página hub é uma boa lista de [links para páginas adequadas à necessidade de informação].
  - E.g., para a consulta [chicago bulls]: a lista do Bob de fontes recomendadas sobre o time Chicago Bulls

HITS – Hyperlink-Induced Topic Search

- Premissa: dois tipos de relevância na web
- Tipo 1: Hubs. Uma página hub é uma boa lista de [links para páginas adequadas à necessidade de informação].
  - E.g., para a consulta [chicago bulls]: a lista do Bob de fontes recomendadas sobre o time Chicago Bulls
- Tipo 2: Autoridades. Uma página de autoridade é uma resposta direta à necessidade de informação

HITS – Hyperlink-Induced Topic Search

- Premissa: dois tipos de relevância na web
- Tipo 1: Hubs. Uma página hub é uma boa lista de [links para páginas adequadas à necessidade de informação].
  - E.g., para a consulta [chicago bulls]: a lista do Bob de fontes recomendadas sobre o time Chicago Bulls
- Tipo 2: Autoridades. Uma página de autoridade é uma resposta direta à necessidade de informação
  - A home page do Chicago Bulls

HITS – Hyperlink-Induced Topic Search

- Premissa: dois tipos de relevância na web
- Tipo 1: Hubs. Uma página hub é uma boa lista de [links para páginas adequadas à necessidade de informação].
  - E.g., para a consulta [chicago bulls]: a lista do Bob de fontes recomendadas sobre o time Chicago Bulls
- Tipo 2: Autoridades. Uma página de autoridade é uma resposta direta à necessidade de informação
  - A home page do Chicago Bulls
  - Por definição: links para páginas de autoridade ocorrem nas páginas hubs

HITS – Hyperlink-Induced Topic Search

- Premissa: dois tipos de relevância na web
- Tipo 1: Hubs. Uma página hub é uma boa lista de [links para páginas adequadas à necessidade de informação].
  - E.g., para a consulta [chicago bulls]: a lista do Bob de fontes recomendadas sobre o time Chicago Bulls
- Tipo 2: Autoridades. Uma página de autoridade é uma resposta direta à necessidade de informação
  - A home page do Chicago Bulls
  - Por definição: links para páginas de autoridade ocorrem nas páginas hubs
- Maior parte das abordagens para busca (incluindo o ranking PageRank) não fazem distinção dos dois tipos de relevância

Hubs e autoridades: Definição

- Uma boa página hub para um tópico linka para muitas páginas de autoridade para aquele tópico

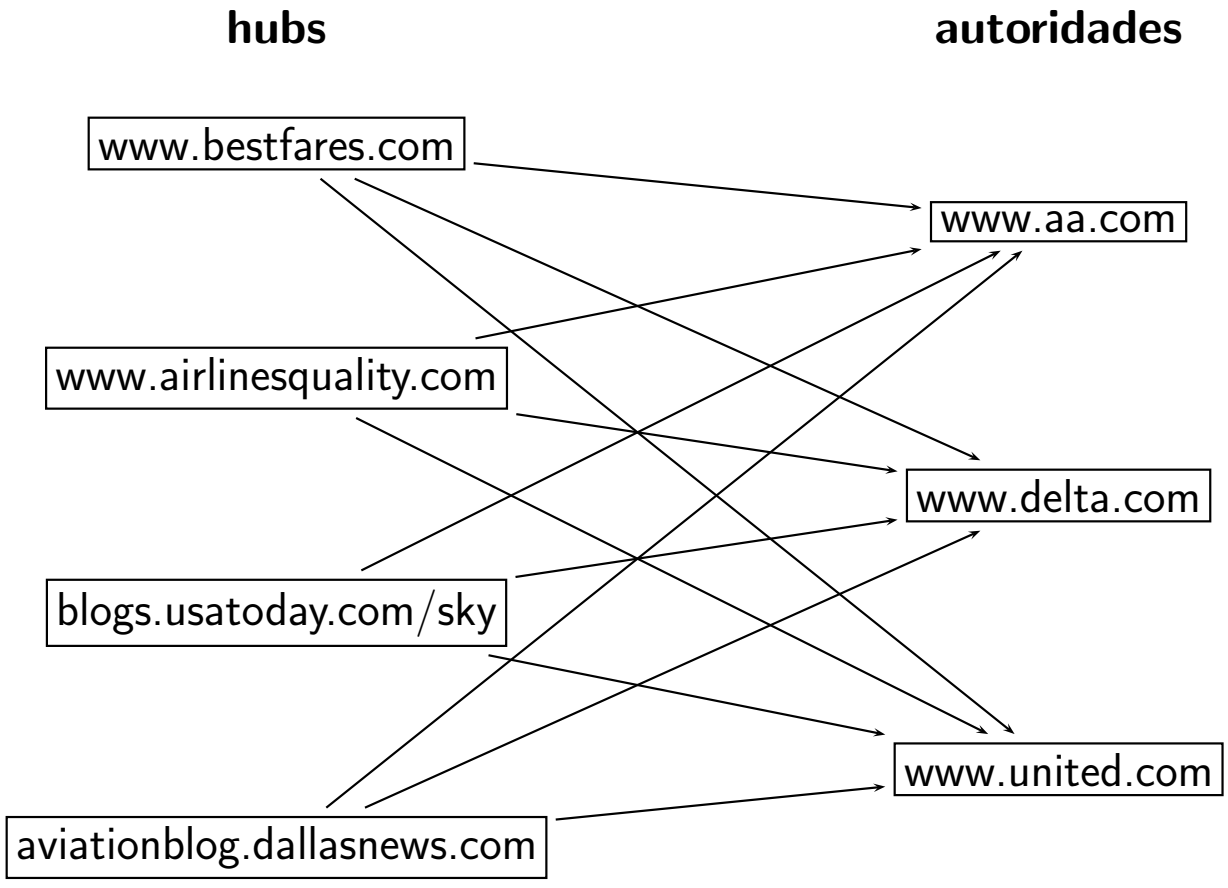
Hubs e autoridades: Definição

- Uma boa página hub para um tópico linka para muitas páginas de autoridade para aquele tópico
- Uma boa página de autoridade para um tópico é linkado por muitas páginas hub para aquele tópico

Hubs e autoridades: Definição

- Uma boa página hub para um tópico linka para muitas páginas de autoridade para aquele tópico
- Uma boa página de autoridade para um tópico é linkado por muitas páginas hub para aquele tópico
- Definição Circular – transformaremos isso em um cálculo iterativo

Exemplo de hubs e autoridades



Como calcular pontuação de páginas de hubs e de autoridades

- Fazer uma busca web normal

Como calcular pontuação de páginas de hubs e de autoridades

- Fazer uma busca web normal
- Chamar o resultado de conjunto raiz

Como calcular pontuação de páginas de hubs e de autoridades

- Fazer uma busca web normal
- Chamar o resultado de conjunto raiz
- Encontrar todas as páginas que são linkadas para essas páginas

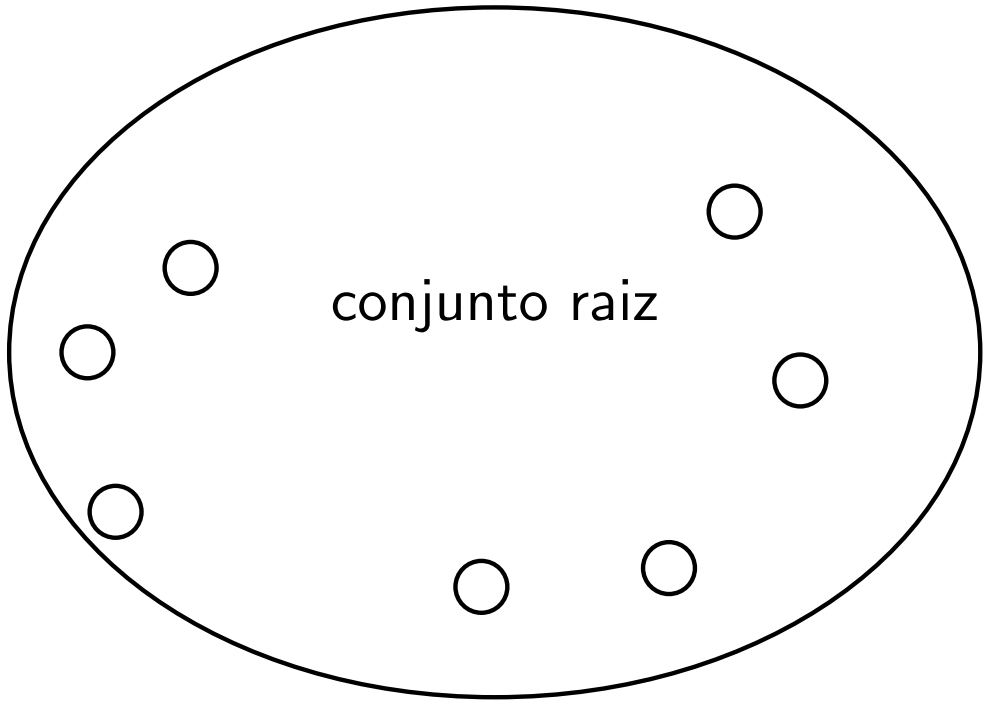
Como calcular pontuação de páginas de hubs e de autoridades

- Fazer uma busca web normal
- Chamar o resultado de conjunto raiz
- Encontrar todas as páginas que são linkadas para essas páginas
- Chamar esse resultado mais amplo de conjunto base

Como calcular pontuação de páginas de hubs e de autoridades

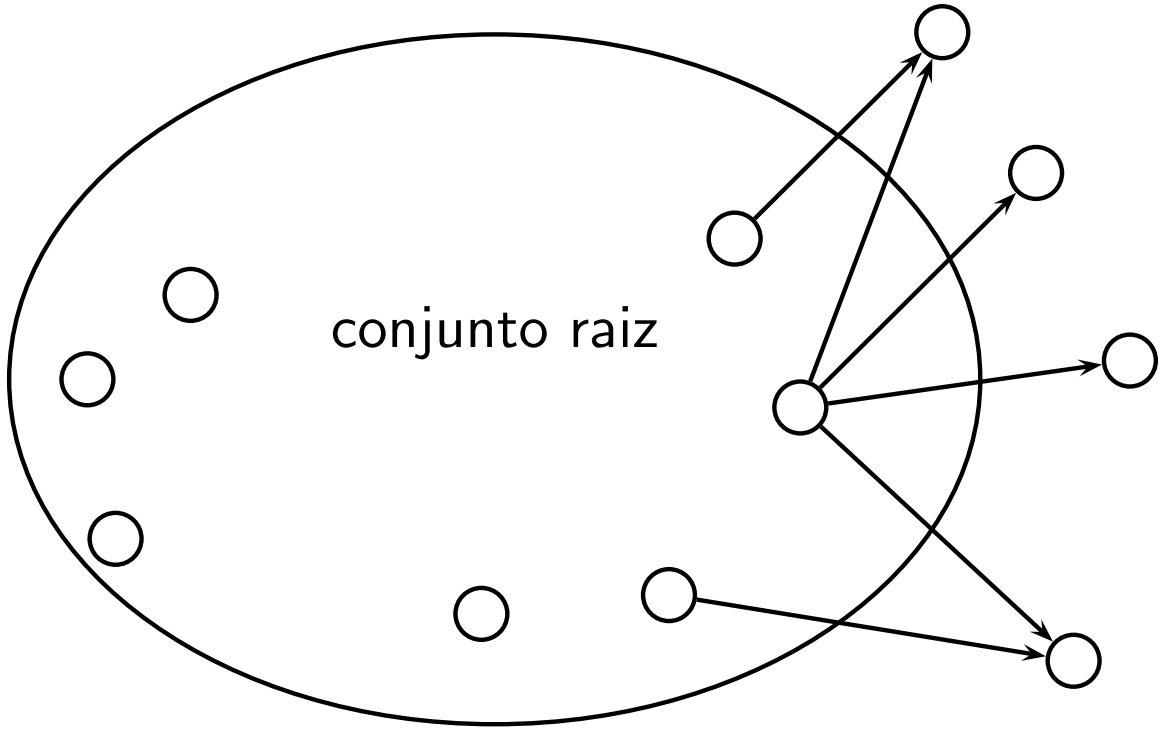
- Fazer uma busca web normal
- Chamar o resultado de conjunto raiz
- Encontrar todas as páginas que são linkadas para essas páginas
- Chamar esse resultado mais amplo de conjunto base
- Finalmente, calcular hubs e autoridades do conjunto base, o qual veremos como um pequeno grafo web

Conjunto raiz e conjunto base (1)



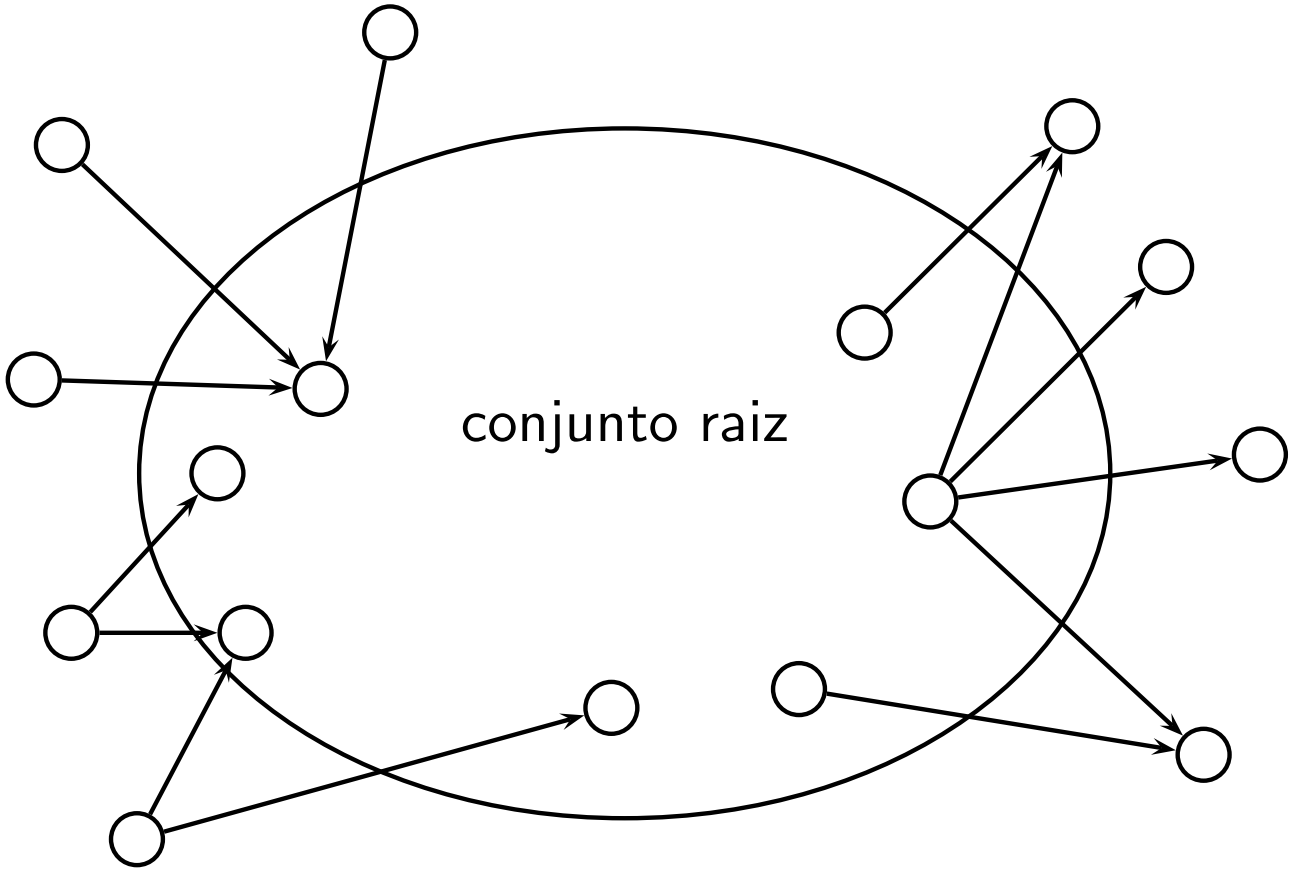
O conjunto raiz

Conjunto raiz e conjunto base (1)



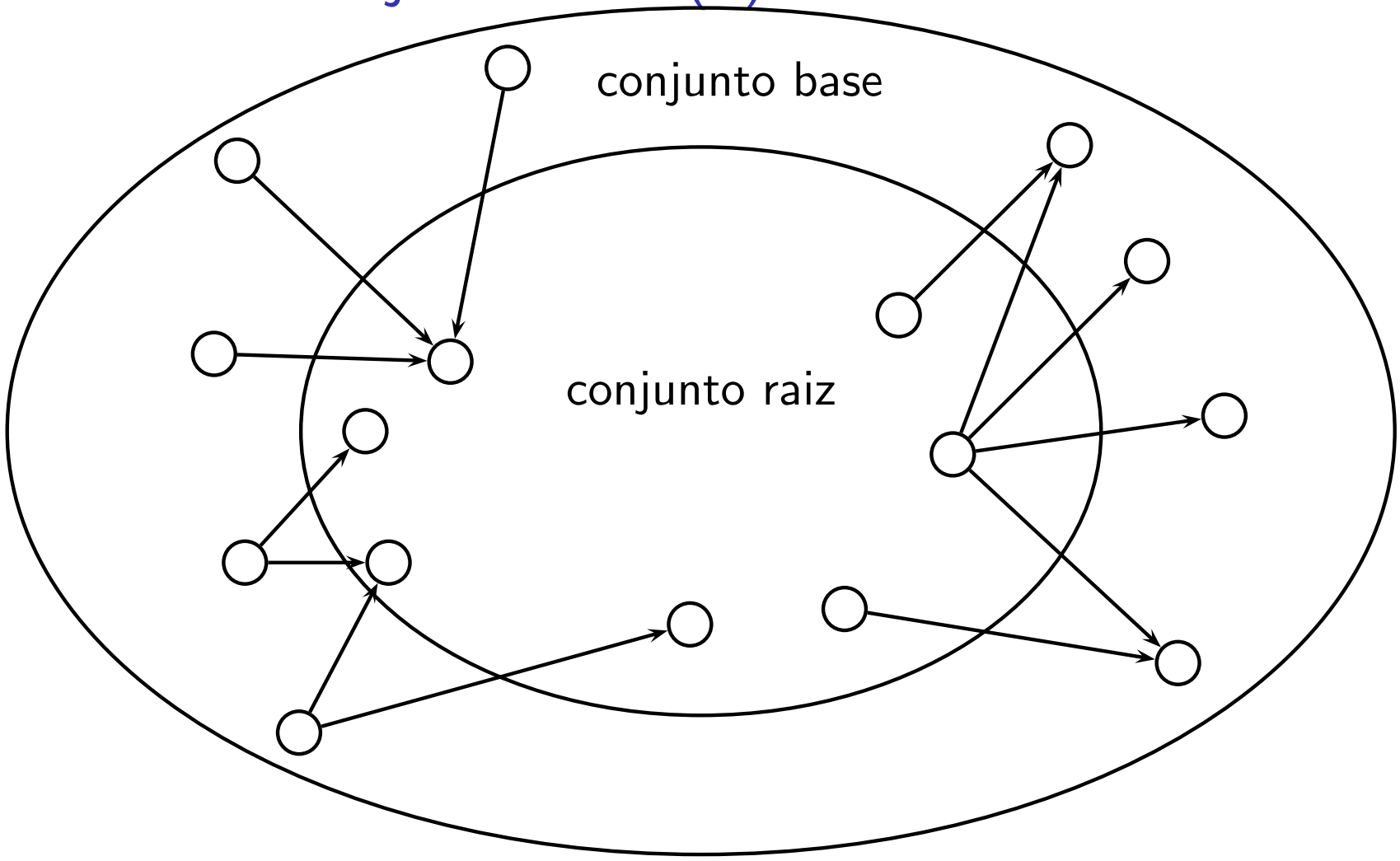
Nós para os quais os nós do conjunto raiz nodes linkam

Conjunto raiz e conjunto base (1)



Nós que linkam para os nós do conjunto raiz

Conjunto raiz e conjunto base (1)



O conjunto base

Conjunto raiz e conjunto base (2)

- Conjunto raiz tipicamente tem 200–1000 nós

Conjunto raiz e conjunto base (2)

- Conjunto raiz tipicamente tem 200–1000 nós
- Conjunto base pode ter até 5000 nós

Conjunto raiz e conjunto base (2)

- Conjunto raiz tipicamente tem 200–1000 nós
- Conjunto base pode ter até 5000 nós
- Cálculo do conjunto base, como mostrado antes:

Conjunto raiz e conjunto base (2)

- Conjunto raiz tipicamente tem 200–1000 nós
- Conjunto base pode ter até 5000 nós
- Cálculo do conjunto base, como mostrado antes:
  - Seguir outlinks processando as páginas no conjunto raiz

Conjunto raiz e conjunto base (2)

- Conjunto raiz tipicamente tem 200–1000 nós
- Conjunto base pode ter até 5000 nós
- Cálculo do conjunto base, como mostrado antes:
  - Seguir outlinks processando as páginas no conjunto raiz
  - Encontrar os inlinks de d ao buscar por todas as páginas contendo um link para d

Conjunto raiz e conjunto base (2)

- Conjunto raiz tipicamente tem 200–1000 nós
- Conjunto base pode ter até 5000 nós
- Cálculo do conjunto base, como mostrado antes:
  - Seguir outlinks processando as páginas no conjunto raiz
  - Encontrar os inlinks de d ao buscar por todas as páginas contendo um link para d
  - Assume existencia de índice que suporta busca por links

Pontuação de Hub e autoridade

- Calcular para cada página d no conjunto base uma pontuação hub $h(d)$ e uma pontuação autoridade $a(d)$

Pontuação de Hub e autoridade

- Calcular para cada página d no conjunto base uma pontuação hub $h(d)$ e uma pontuação autoridade $a(d)$
- Inicialização: para todos d: $h(d) = 1$, $a(d) = 1$

Pontuação de Hub e autoridade

- Calcular para cada página d no conjunto base uma pontuação hub $h(d)$ e uma pontuação autoridade $a(d)$
- Inicialização: para todos d: $h(d) = 1$, $a(d) = 1$
- Iterativamento atualizar todos $h(d), a(d)$

Pontuação de Hub e autoridade

- Calcular para cada página d no conjunto base uma pontuação hub $h(d)$ e uma pontuação autoridade $a(d)$
- Inicialização: para todos d: $h(d) = 1$, $a(d) = 1$
- Iterativamento atualizar todos $h(d), a(d)$
- Após convergência:

Pontuação de Hub e autoridade

- Calcular para cada página d no conjunto base uma pontuação hub $h(d)$ e uma pontuação autoridade $a(d)$
- Inicialização: para todos d: $h(d) = 1$, $a(d) = 1$
- Iterativamento atualizar todos $h(d), a(d)$
- Após convergência:
  - Saída: páginas com maiores pontuações de h como as páginas top hubs

Pontuação de Hub e autoridade

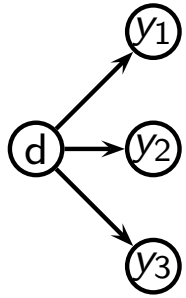
- Calcular para cada página d no conjunto base uma pontuação hub $h(d)$ e uma pontuação autoridade $a(d)$
- Inicialização: para todos d: $h(d) = 1$, $a(d) = 1$
- Iterativamento atualizar todos $h(d), a(d)$
- Após convergência:
  - Saída: páginas com maiores pontuações de h como as páginas top hubs
  - Saída: páginas com maiores pontuações a como as páginas top de autoridades

Pontuação de Hub e autoridade

- Calcular para cada página d no conjunto base uma pontuação hub $h(d)$ e uma pontuação autoridade $a(d)$
- Inicialização: para todos d: $h(d) = 1$, $a(d) = 1$
- Iterativamento atualizar todos $h(d), a(d)$
- Após convergência:
  - Saída: páginas com maiores pontuações de h como as páginas top hubs
  - Saída: páginas com maiores pontuações a como as páginas top de autoridades
  - Portanto, produzimos duas listas ordenadas

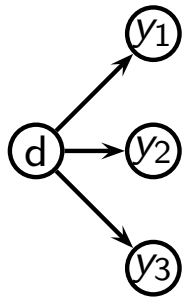
Atualização iterativa

- Para todos d: $h(d) = \sum_{d \mapsto y} a(y)$

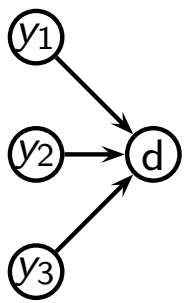


Atualização iterativa

- Para todos d: $h(d) = \sum_{d \mapsto y} a(y)$

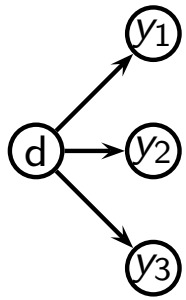


- Para todos d: $a(d) = \sum_{y \mapsto d} h(y)$

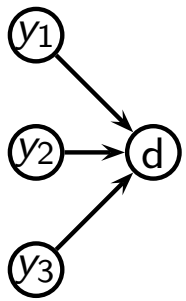


Atualização iterativa

- Para todos d: $h(d) = \sum_{d \mapsto y} a(y)$



- Para todos d: $a(d) = \sum_{y \mapsto d} h(y)$



- Iterar esses dois passos até convergência

Detalhes

- Mudança de escala

Detalhes

- Mudança de escala
  - Prevenir que os valores de $a()$ e $h()$ fiquem muito altos, reduzir escala após cada iteração

Detalhes

- Mudança de escala
  - Prevenir que os valores de $a()$ e $h()$ fiquem muito altos, reduzir escala após cada iteração
  - Valor do fator de escala não altera os resultados

Detalhes

- Mudança de escala
  - Prevenir que os valores de $a()$ e $h()$ fiquem muito altos, reduzir escala após cada iteração
  - Valor do fator de escala não altera os resultados
  - Mais importante o relativo em vez dos valores absolutos da pontuação

Detalhes

- Mudança de escala
  - Prevenir que os valores de $a()$ e $h()$ fiquem muito altos, reduzir escala após cada iteração
  - Valor do fator de escala não altera os resultados
  - Mais importante o relativo em vez dos valores absolutos da pontuação
- Na maior parte dos casos, o algoritmo converge após algumas iterações

Autoridades para consulta [Chicago Bulls]

0.85 www.nba.com/bulls

0.25 www.essex1.com/people/jmiller/bulls.htm
"da Bulls"

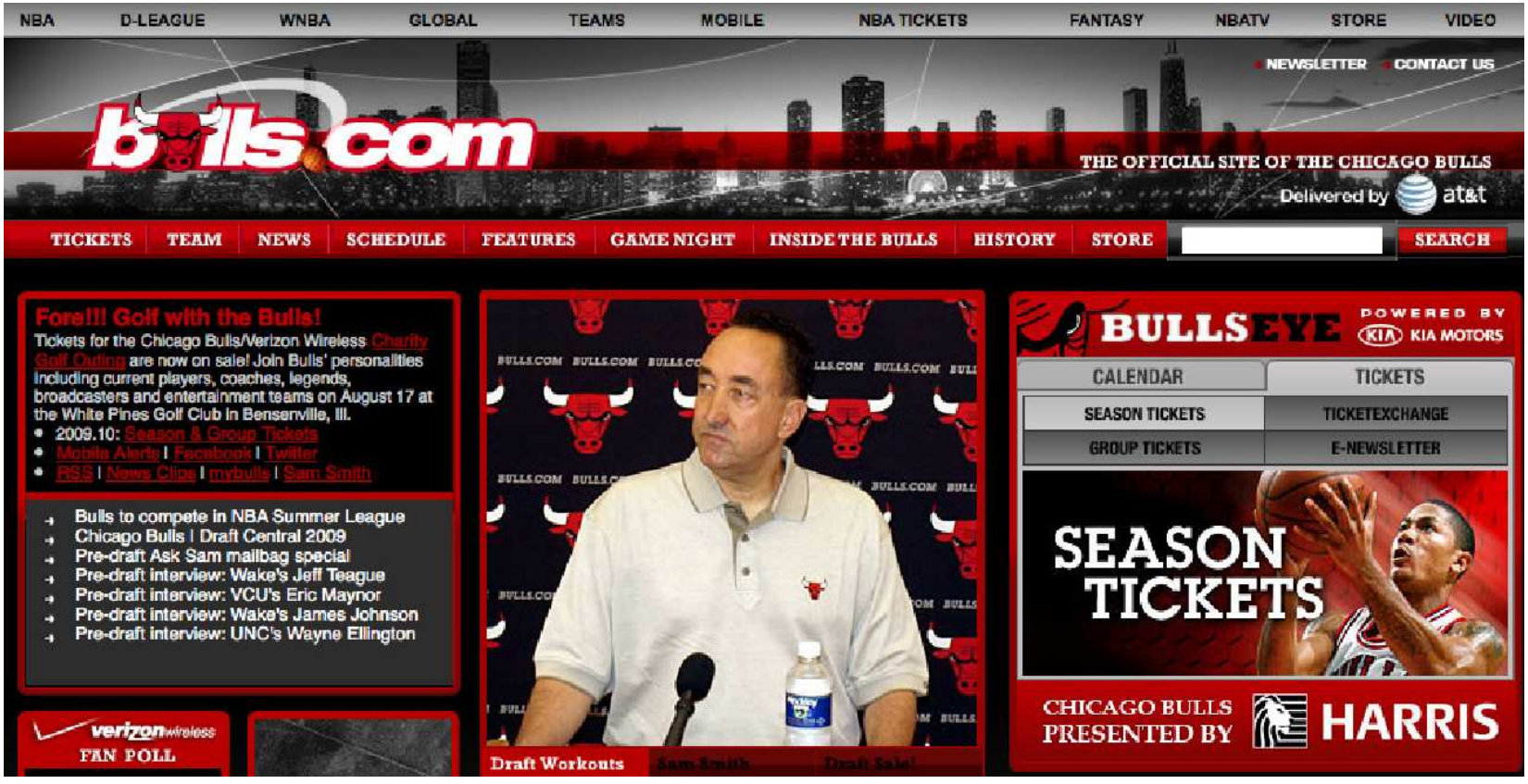
0.20 www.nando.net/SportServer/basketball/nba/chi.html
"The Chicago Bulls"

0.15 users.aol.com/rynocub/bulls.htm
"The Chicago Bulls Home Page"

0.13 www.geocities.com/Colosseum/6095
"Chicago Bulls"

(Ben-Shaul et al, WWW8)

A página de autoridade para [Chicago Bulls]



Hubs para [Chicago Bulls]

1.62 www.geocities.com/Colosseum/1778
"Unbelieveabulls!!!!!"

1.24 www.webring.org/cgi-bin/webring?ring=chbulls
"Erin's Chicago Bulls Page"

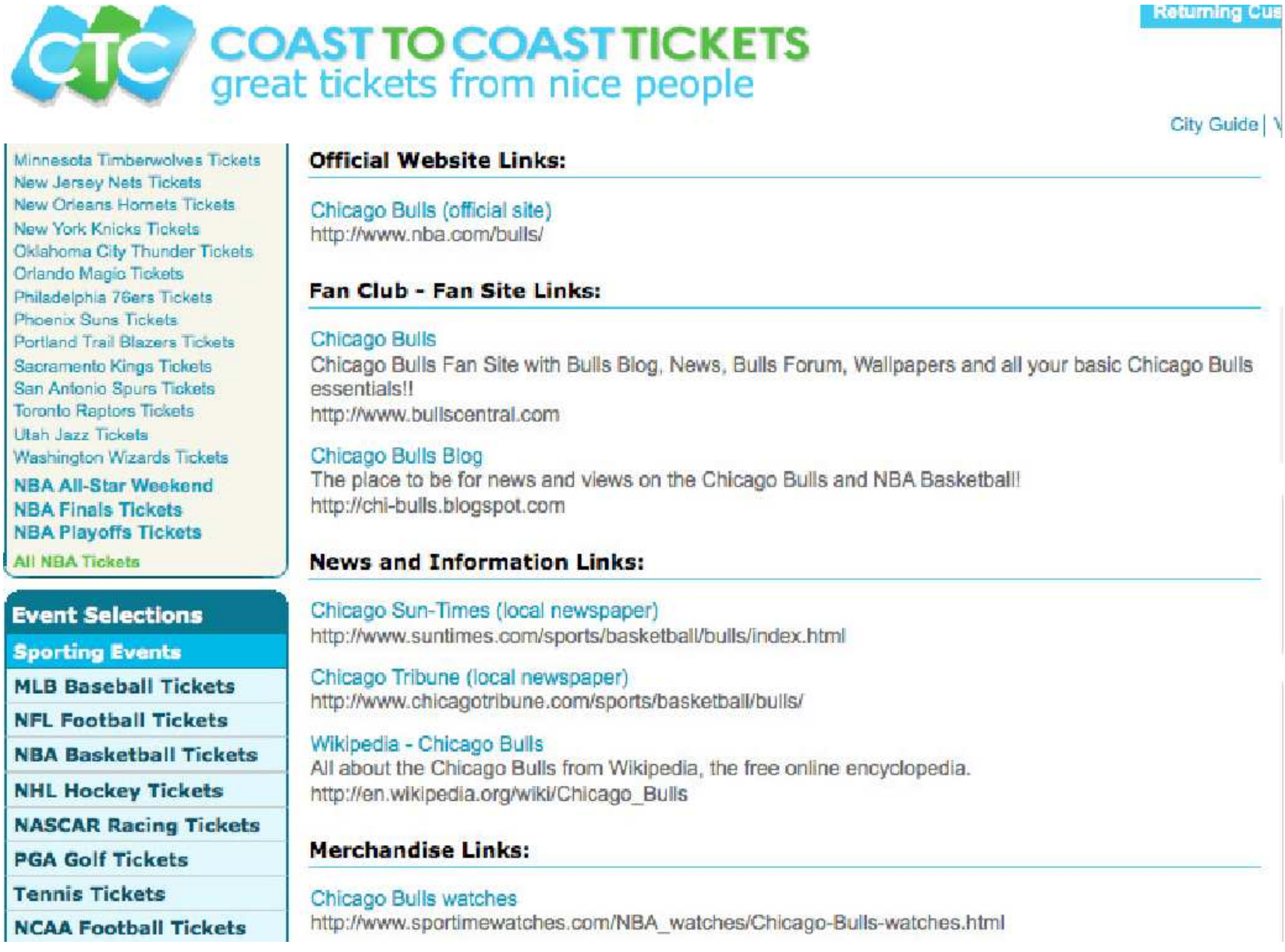
0.74 www.geocities.com/Hollywood/Lot/3330/Bulls.html
"Chicago Bulls"

0.52 www.nobull.net/web_position/kw-search-15-M2.htm
"Excite Search Results: bulls"

0.52 www.halcyon.com/wordsltd/bball/bulls.htm
"Chicago Bulls Links"

(Ben-Shaul et al, WWW8)

Um hub para [Chicago Bulls]



Hubs & Authorities: Comentários

- HITS pode juntar páginas boas independentemente do conteúdo da página

Hubs & Authorities: Comentários

- HITS pode juntar páginas boas independentemente do conteúdo da página
- Uma vez que o conjunto base é construído, fazemos a análise de links, sem fazer verificação do texto

Hubs & Authorities: Comentários

- HITS pode juntar páginas boas independentemente do conteúdo da página
- Uma vez que o conjunto base é construído, fazemos a análise de links, sem fazer verificação do texto
- Páginas no conjunto base costumam não ter nenhuma dos termos de consulta

Hubs & Authorities: Comentários

- HITS pode juntar páginas boas independentemente do conteúdo da página
- Uma vez que o conjunto base é construído, fazemos a análise de links, sem fazer verificação do texto
- Páginas no conjunto base costumam não ter nenhuma dos termos de consulta
- Em teoria, uma consulta em inglês pode recuperar páginas em japonês

Hubs & Authorities: Comentários

- HITS pode juntar páginas boas independentemente do conteúdo da página
- Uma vez que o conjunto base é construído, fazemos a análise de links, sem fazer verificação do texto
- Páginas no conjunto base costumam não ter nenhuma dos termos de consulta
- Em teoria, uma consulta em inglês pode recuperar páginas em japonês
  - Isso acontece se na estrutura de links houver ligação entre páginas dos dois idiomas

Hubs & Authorities: Comentários

- HITS pode juntar páginas boas independentemente do conteúdo da página
- Uma vez que o conjunto base é construído, fazemos a análise de links, sem fazer verificação do texto
- Páginas no conjunto base costumam não ter nenhuma dos termos de consulta
- Em teoria, uma consulta em inglês pode recuperar páginas em japonês
  - Isso acontece se na estrutura de links houver ligação entre páginas dos dois idiomas
- Risco: desvio do tópico – as páginas encontradas por seguir links podem não estar relacionadas à consulta original

Sobre a convergência

Sobre a convergência

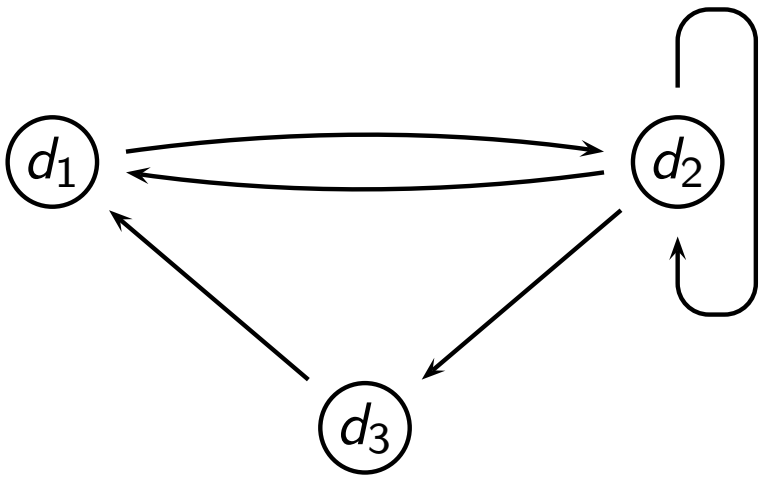
- Definimos uma matriz de adjacência A com $N \times N$ elementos. (Antes chamamos essa matriz de "matriz de links") .

Sobre a convergência

- Definimos uma matriz de adjacência A com $N \times N$ elementos. (Antes chamamos essa matriz de "matriz de links") .
- Para $1 \leq i, j \leq N$, o valor da matriz A_{ij} nos diz se há um link da página i para página j ($A_{ij} = 1$) ou não ($A_{ij} = 0$).

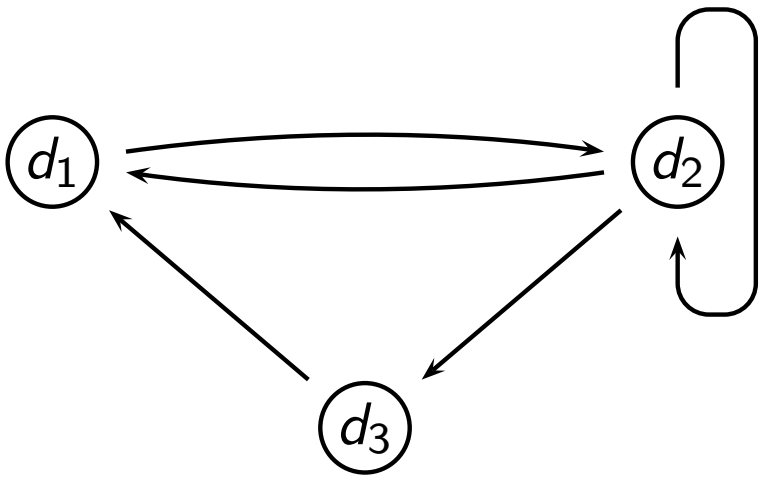
Sobre a convergência

- Definimos uma matriz de adjacência A com $N \times N$ elementos. (Antes chamamos essa matriz de "matriz de links") .
- Para $1 \leq i, j \leq N$, o valor da matriz A_{ij} nos diz se há um link da página i para página j ($A_{ij} = 1$) ou não ($A_{ij} = 0$).
- Exemplo:



Sobre a convergência

- Definimos uma matriz de adjacência A com $N \times N$ elementos. (Antes chamamos essa matriz de "matriz de links") .
- Para $1 \leq i, j \leq N$, o valor da matriz A_{ij} nos diz se há um link da página i para página j ($A_{ij} = 1$) ou não ($A_{ij} = 0$).
- Exemplo:



	d_1	d_2	d_3
d_1	0	1	0
d_2	1	1	1
d_3	1	0	0

Escrever regras de atualização com operações matriciais

- Definir o vetor de pontuações hub $\vec{h} = (h_1, \ldots, h_N)$. h_i é a pontuação hub da página d_i.

Escrever regras de atualização com operações matriciais

- Definir o vetor de pontuações hub $\vec{h} = (h_1, \ldots, h_N)$. h_i é a pontuação hub da página d_i.
- De maneira similar para $\vec{a}$, o vetor de pontuação de autoridade

Escrever regras de atualização com operações matriciais

- Definir o vetor de pontuações hub $\vec{h} = (h_1, \ldots, h_N)$. h_i é a pontuação hub da página d_i.
- De maneira similar para $\vec{a}$, o vetor de pontuação de autoridade
- Agora podemos escrever $h(d) = \sum_{d \mapsto y} a(y)$ como uma operação matricial: $\vec{h} = A\vec{a}$ ...

Escrever regras de atualização com operações matriciais

- Definir o vetor de pontuações hub $\vec{h} = (h_1, \ldots, h_N)$. h_i é a pontuação hub da página d_i.
- De maneira similar para $\vec{a}$, o vetor de pontuação de autoridade
- Agora podemos escrever $h(d) = \sum_{d \mapsto y} a(y)$ como uma operação matricial: $\vec{h} = A\vec{a}$ ...
- ... e escrevemos $a(d) = \sum_{y \mapsto d} h(y)$ como $\vec{a} = A^T \vec{h}$

Escrever regras de atualização com operações matriciais

- Definir o vetor de pontuações hub $\vec{h} = (h_1, \ldots, h_N)$. h_i é a pontuação hub da página d_i.
- De maneira similar para $\vec{a}$, o vetor de pontuação de autoridade
- Agora podemos escrever $h(d) = \sum_{d \mapsto y} a(y)$ como uma operação matricial: $\vec{h} = A\vec{a}$ ...
- ... e escrevemos $a(d) = \sum_{y \mapsto d} h(y)$ como $\vec{a} = A^T \vec{h}$
- HITS algoritmo em notação matricial:

Escrever regras de atualização com operações matriciais

- Definir o vetor de pontuações hub $\vec{h} = (h_1, \ldots, h_N)$. h_i é a pontuação hub da página d_i.
- De maneira similar para $\vec{a}$, o vetor de pontuação de autoridade
- Agora podemos escrever $h(d) = \sum_{d \mapsto y} a(y)$ como uma operação matricial: $\vec{h} = A\vec{a}$ ...
- ... e escrevemos $a(d) = \sum_{y \mapsto d} h(y)$ como $\vec{a} = A^T \vec{h}$
- HITS algoritmo em notação matricial:
  - Calcular $\vec{h} = A\vec{a}$

Escrever regras de atualização com operações matriciais

- Definir o vetor de pontuações hub $\vec{h} = (h_1, \ldots, h_N)$. h_i é a pontuação hub da página d_i.
- De maneira similar para $\vec{a}$, o vetor de pontuação de autoridade
- Agora podemos escrever $h(d) = \sum_{d \mapsto y} a(y)$ como uma operação matricial: $\vec{h} = A\vec{a}$ ...
- ... e escrevemos $a(d) = \sum_{y \mapsto d} h(y)$ como $\vec{a} = A^T \vec{h}$
- HITS algoritmo em notação matricial:
  - Calcular $\vec{h} = A\vec{a}$
  - Calcular $\vec{a} = A^T \vec{h}$

Escrever regras de atualização com operações matriciais

- Definir o vetor de pontuações hub $\vec{h} = (h_1, \ldots, h_N)$. h_i é a pontuação hub da página d_i.
- De maneira similar para $\vec{a}$, o vetor de pontuação de autoridade
- Agora podemos escrever $h(d) = \sum_{d \mapsto y} a(y)$ como uma operação matricial: $\vec{h} = A\vec{a}$ ...
- ... e escrevemos $a(d) = \sum_{y \mapsto d} h(y)$ como $\vec{a} = A^T \vec{h}$
- HITS algoritmo em notação matricial:
  - Calcular $\vec{h} = A\vec{a}$
  - Calcular $\vec{a} = A^T \vec{h}$
  - Iterar até convergência

HITS como um problema de auto-vetor

- Algoritmo HITS em notação matricial. Iterar:
  - Calcular $\vec{h} = A\vec{a}$
  - Calcular $\vec{a} = A^T\vec{h}$

HITS como um problema de auto-vetor

- Algoritmo HITS em notação matricial. Iterar:
  - Calcular $\vec{h} = A\vec{a}$
  - Calcular $\vec{a} = A^T\vec{h}$
- Por substituição temos: $\vec{h} = AA^T\vec{h}$ e $\vec{a} = A^TA\vec{a}$

HITS como um problema de auto-vetor

- Algoritmo HITS em notação matricial. Iterar:
  - Calcular $\vec{h} = A\vec{a}$
  - Calcular $\vec{a} = A^T\vec{h}$
- Por substituição temos: $\vec{h} = AA^T\vec{h}$ e $\vec{a} = A^TA\vec{a}$
- Então, $\vec{h}$ é um auto-vetor de AA^T e $\vec{a}$ é um auto-vetor de A^TA.

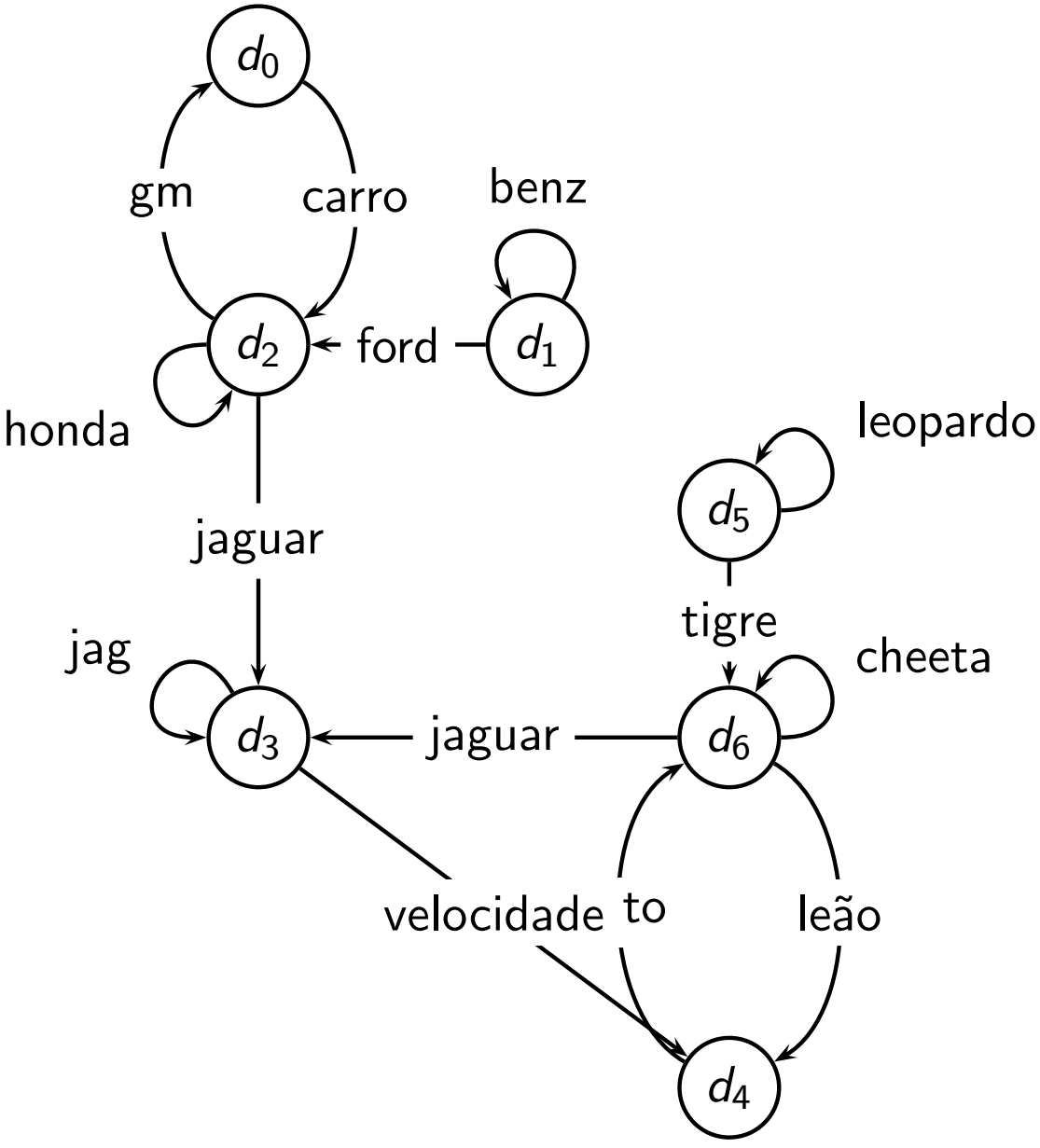
HITS como um problema de auto-vetor

- Algoritmo HITS em notação matricial. Iterar:
  - Calcular $\vec{h} = A\vec{a}$
  - Calcular $\vec{a} = A^T\vec{h}$
- Por substituição temos: $\vec{h} = AA^T\vec{h}$ e $\vec{a} = A^TA\vec{a}$
- Então, $\vec{h}$ é um auto-vetor de AA^T e $\vec{a}$ é um auto-vetor de A^TA.
- Então o algoritmo HITS é na verdade um caso especial do método da potência e pontuações hub e autoridade são auto-vetores

HITS como um problema de auto-vetor

- Algoritmo HITS em notação matricial. Iterar:
  - Calcular $\vec{h} = A\vec{a}$
  - Calcular $\vec{a} = A^T\vec{h}$
- Por substituição temos: $\vec{h} = AA^T\vec{h}$ e $\vec{a} = A^TA\vec{a}$
- Então, $\vec{h}$ é um auto-vetor de AA^T e $\vec{a}$ é um auto-vetor de A^TA.
- Então o algoritmo HITS é na verdade um caso especial do método da potência e pontuações hub e autoridade são auto-vetores
- HITS e PageRank ambos formalizam a análise de links como o problema de encontrar um auto-vetor.

Exemplo grafo web



Matriz pura A para HITS

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0	0	1	0	0	0	0
d_1	0	1	1	0	0	0	0
d_2	1	0	1	2	0	0	0
d_3	0	0	0	1	1	0	0
d_4	0	0	0	0	0	0	1
d_5	0	0	0	0	0	1	1
d_6	0	0	0	2	1	0	1

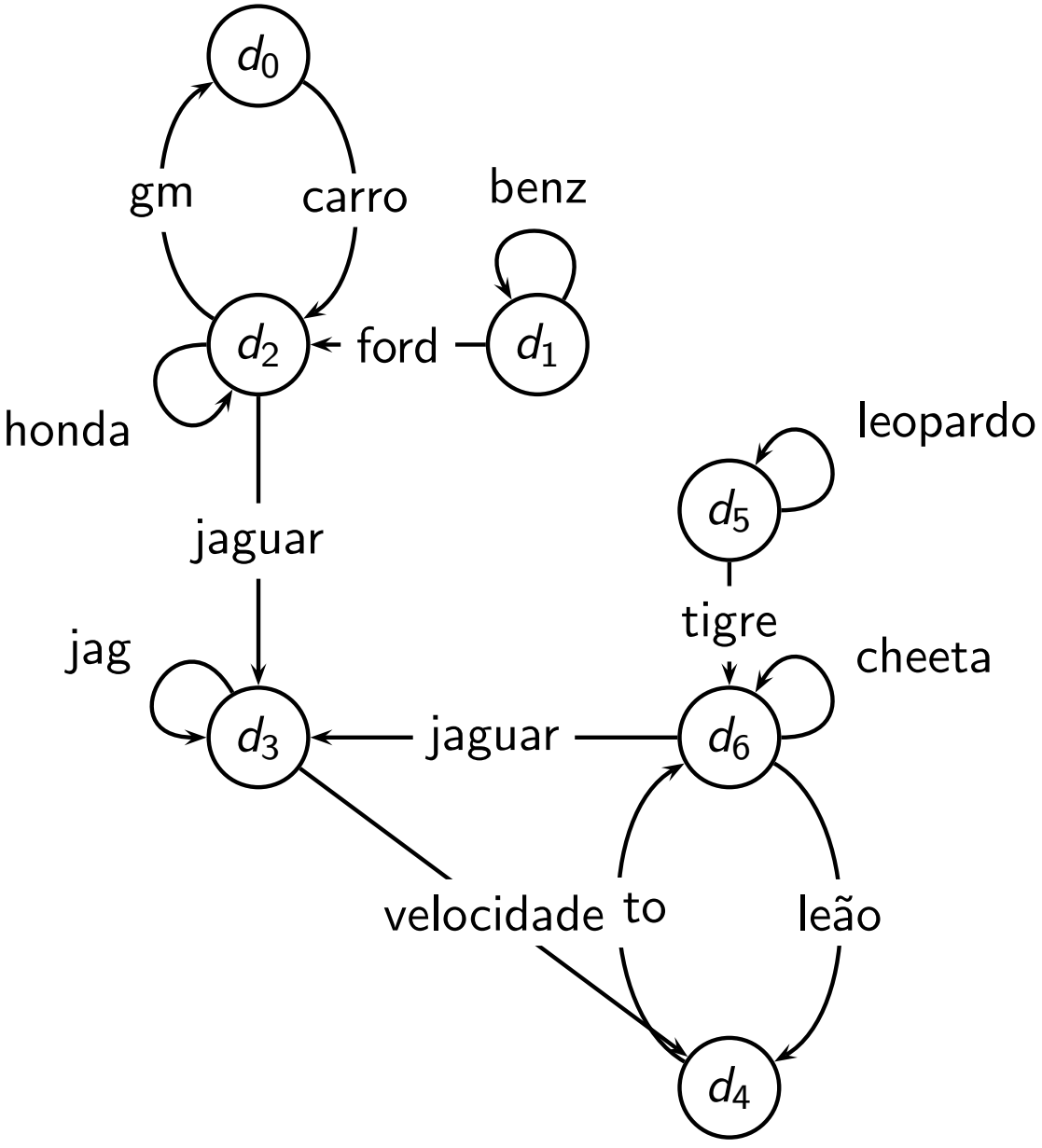
Vetores de Hub $h_0, \vec{h_i} = \frac{1}{d_i} A \cdot \vec{a_i}, i \geq 1$

	$\vec{h}_0$	$\vec{h}_1$	$\vec{h}_2$	$\vec{h}_3$	$\vec{h}_4$	$\vec{h}_5$
d_0	0.14	0.06	0.04	0.04	0.03	0.03
d_1	0.14	0.08	0.05	0.04	0.04	0.04
d_2	0.14	0.28	0.32	0.33	0.33	0.33
d_3	0.14	0.14	0.17	0.18	0.18	0.18
d_4	0.14	0.06	0.04	0.04	0.04	0.04
d_5	0.14	0.08	0.05	0.04	0.04	0.04
d_6	0.14	0.30	0.33	0.34	0.35	0.35

Vetores de autoridade $\vec{a_i} = \frac{1}{c_i} A^T \cdot \vec{h}_{i-1}, i \geq 1$

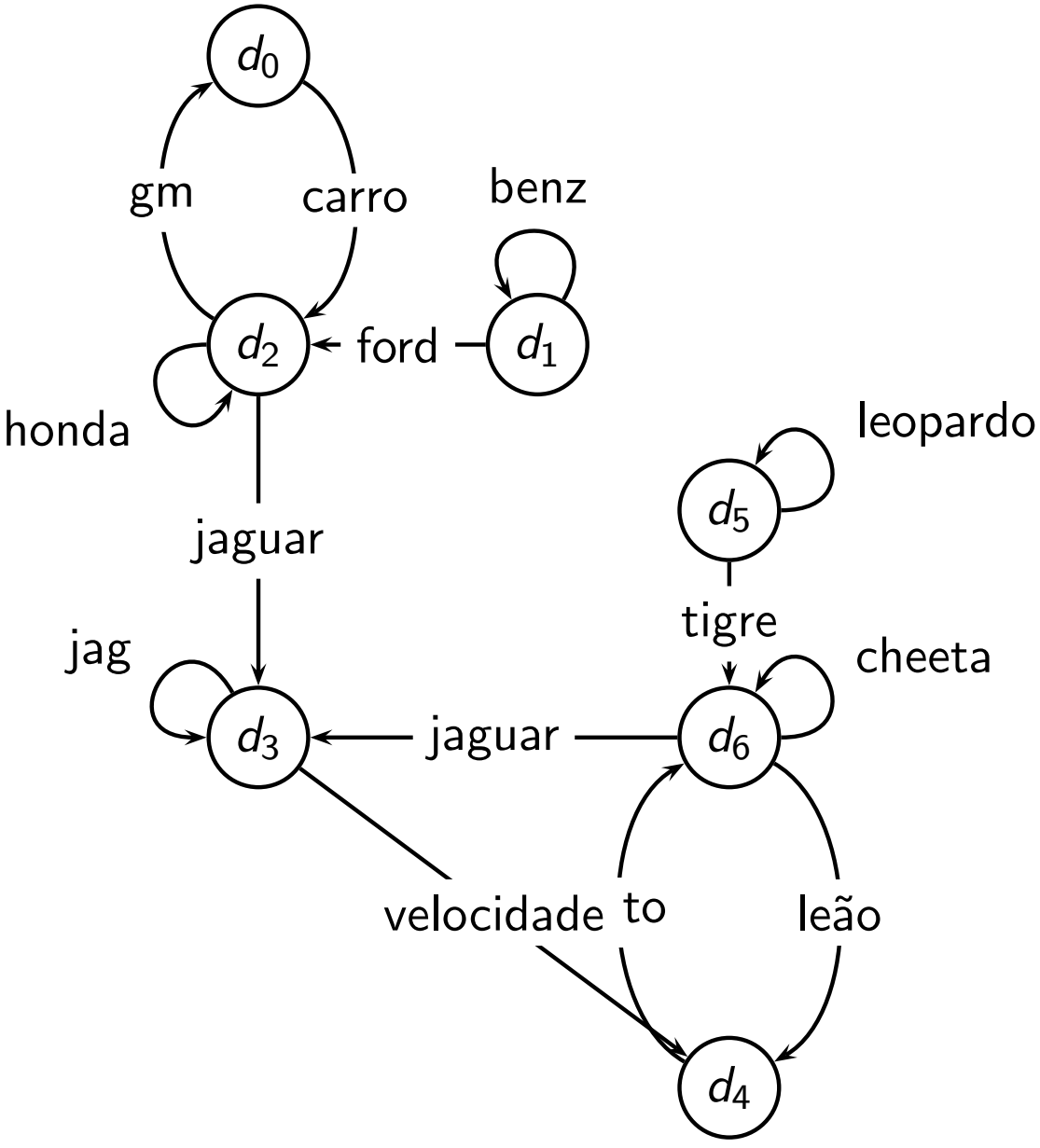
	$\vec{a}_1$	$\vec{a}_2$	$\vec{a}_3$	$\vec{a}_4$	$\vec{a}_5$	$\vec{a}_6$	$\vec{a}_7$
d_0	0.06	0.09	0.10	0.10	0.10	0.10	0.10
d_1	0.06	0.03	0.01	0.01	0.01	0.01	0.01
d_2	0.19	0.14	0.13	0.12	0.12	0.12	0.12
d_3	0.31	0.43	0.46	0.46	0.46	0.47	0.47
d_4	0.13	0.14	0.16	0.16	0.16	0.16	0.16
d_5	0.06	0.03	0.02	0.01	0.01	0.01	0.01
d_6	0.19	0.14	0.13	0.13	0.13	0.13	0.13

Exemplo grafo web



	a	h
d_0	0.10	0.03
d_1	0.01	0.04
d_2	0.12	0.33
d_3	0.47	0.18
d_4	0.16	0.04
d_5	0.01	0.04
d_6	0.13	0.35

Exemplo grafo web



Páginas com maior grau de entrada: d_2, d_3, d_6

Páginas com maior grau de saída: d_2, d_6

Páginas com maior PageRank: d_6

Páginas com maior pontuação de hub: d_6 (similar: d_2)

Páginas com maior pontuação de autoridade: d_3

PageRank vs. HITS

- PageRank pode ser pré-calculado, HITS tem que ser calcular em tempo da consulta

PageRank vs. HITS

- PageRank pode ser pré-calculado, HITS tem que ser calcular em tempo da consulta
  - HITS é muito caro para a maior parte das aplicações

PageRank vs. HITS

- PageRank pode ser pré-calculado, HITS tem que ser calcular em tempo da consulta
  - HITS é muito caro para a maior parte das aplicações
- PageRank e HITS fazem duas escolhas de projetos em relação (i) a formalização do problema de auto-vetor/valor (ii) o conjunto de páginas para aplicar a formalização

PageRank vs. HITS

- PageRank pode ser pré-calculado, HITS tem que ser calcular em tempo da consulta
  - HITS é muito caro para a maior parte das aplicações
- PageRank e HITS fazem duas escolhas de projetos em relação (i) a formalização do problema de auto-vetor/valor (ii) o conjunto de páginas para aplicar a formalização
- Esses dois são ortogonais

PageRank vs. HITS

- PageRank pode ser pré-calculado, HITS tem que ser calcular em tempo da consulta
  - HITS é muito caro para a maior parte das aplicações
- PageRank e HITS fazem duas escolhas de projetos em relação (i) a formalização do problema de auto-vetor/valor (ii) o conjunto de páginas para aplicar a formalização
- Esses dois são ortogonais
  - Poderíamos também aplicar HITS para a web inteira e PageRank para um pequeno conjunto base

PageRank vs. HITS

- PageRank pode ser pré-calculado, HITS tem que ser calcular em tempo da consulta
  - HITS é muito caro para a maior parte das aplicações
- PageRank e HITS fazem duas escolhas de projetos em relação (i) a formalização do problema de auto-vetor/valor (ii) o conjunto de páginas para aplicar a formalização
- Esses dois são ortogonais
  - Poderíamos também aplicar HITS para a web inteira e PageRank para um pequeno conjunto base
- Empírico: na web, um bom hub quase sempre é uma boa autoridade

PageRank vs. HITS

- PageRank pode ser pré-calculado, HITS tem que ser calcular em tempo da consulta
  - HITS é muito caro para a maior parte das aplicações
- PageRank e HITS fazem duas escolhas de projetos em relação (i) a formalização do problema de auto-vetor/valor (ii) o conjunto de páginas para aplicar a formalização
- Esses dois são ortogonais
  - Poderíamos também aplicar HITS para a web inteira e PageRank para um pequeno conjunto base
- Empírico: na web, um bom hub quase sempre é uma boa autoridade
- A diferença real do ranking usando PageRank e HITS não é tão grande quanto se poderia esperar