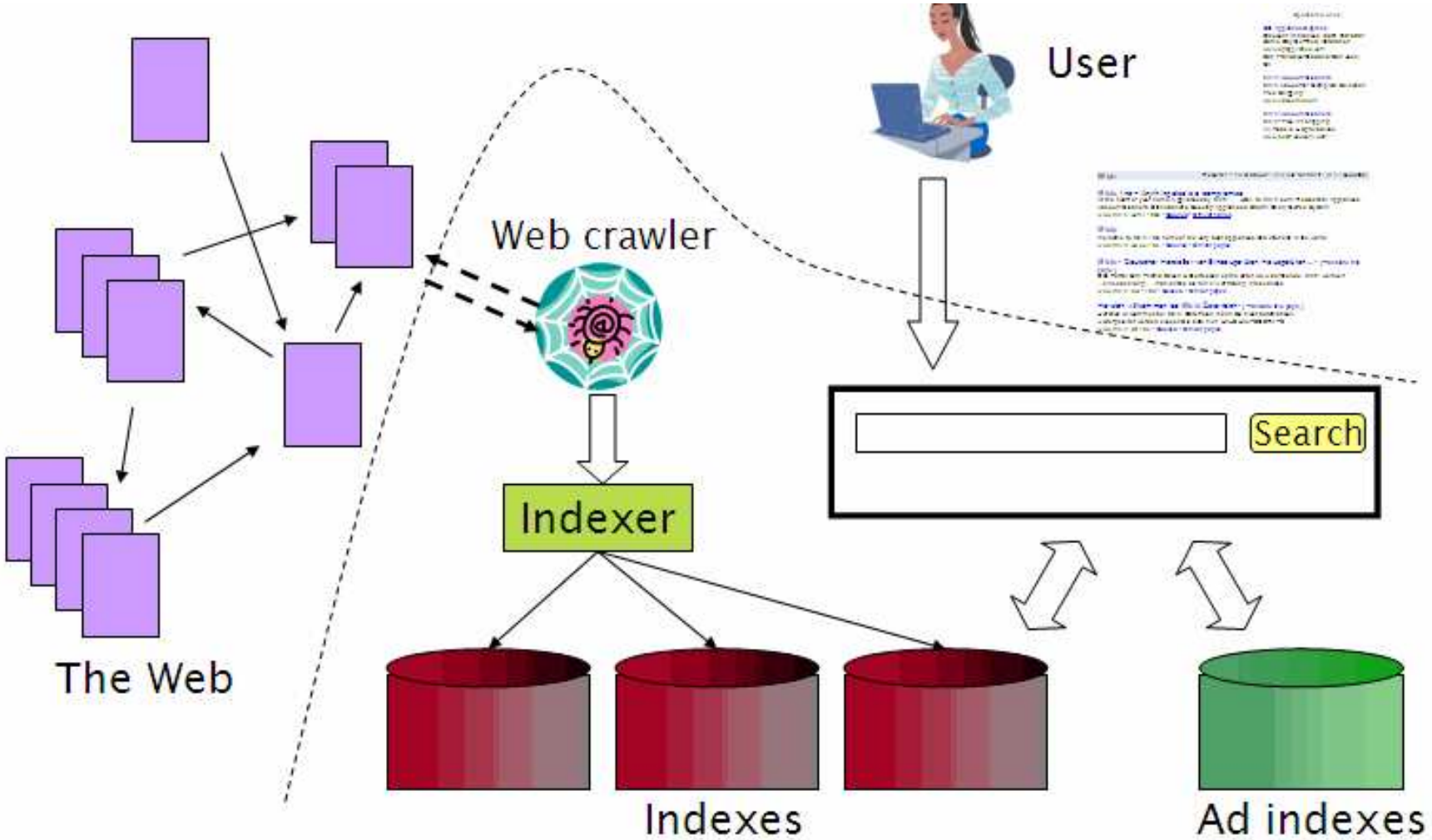
Organização e Recuperação de Informação: Busca web

Marcelo K. A.

Faculdade de Computação - UFU

Conteúdo

Visão geral sobre busca na Web



Sem buscadores, a web não funcionaria

- Sem busca, conteúdo é difícil de achar.

Sem buscadores, a web não funcionaria

- Sem busca, conteúdo é difícil de achar.
- $\rightarrow$ Sem busca, não há incentivo em criar conteúdo.

Sem buscadores, a web não funcionaria

- Sem busca, conteúdo é difícil de achar.
- $\rightarrow$ Sem busca, não há incentivo em criar conteúdo.
  - Porque publicar algo se ninguém for ler?

Sem buscadores, a web não funcionaria

- Sem busca, conteúdo é difícil de achar.
- $\rightarrow$ Sem busca, não há incentivo em criar conteúdo.
  - Porque publicar algo se ninguém for ler?
  - Porque publicar se não recebo receita de anúncios disso?

Sem buscadores, a web não funcionaria

- Sem busca, conteúdo é difícil de achar.
- $\rightarrow$ Sem busca, não há incentivo em criar conteúdo.
  - Porque publicar algo se ninguém for ler?
  - Porque publicar se não recebo receita de anúncios disso?
- Alguém precisa pagar pela web

Sem buscadores, a web não funcionaria

- Sem busca, conteúdo é difícil de achar.
- $\rightarrow$ Sem busca, não há incentivo em criar conteúdo.
  - Porque publicar algo se ninguém for ler?
  - Porque publicar se não recebo receita de anúncios disso?
- Alguém precisa pagar pela web
  - Servidores, infraestrutura web, criação de conteúdo

Sem buscadores, a web não funcionaria

- Sem busca, conteúdo é difícil de achar.
- $\rightarrow$ Sem busca, não há incentivo em criar conteúdo.
  - Porque publicar algo se ninguém for ler?
  - Porque publicar se não recebo receita de anúncios disso?
- Alguém precisa pagar pela web
  - Servidores, infraestrutura web, criação de conteúdo
  - Grande parte hoje é paga por anúncios.

Sem buscadores, a web não funcionaria

- Sem busca, conteúdo é difícil de achar.
- $\rightarrow$ Sem busca, não há incentivo em criar conteúdo.
  - Porque publicar algo se ninguém for ler?
  - Porque publicar se não recebo receita de anúncios disso?
- Alguém precisa pagar pela web
  - Servidores, infraestrutura web, criação de conteúdo
  - Grande parte hoje é paga por anúncios.
  - A busca paga a web.

Agregação de interesses

- Característica única da web: um número de pessoas geograficamente dispersas com interesses similares podem se comunicar

Agregação de interesses

- Característica única da web: um número de pessoas geograficamente dispersas com interesses similares podem se comunicar
  - Comunidade dos pais de crianças com hemofilia

Agregação de interesses

- Característica única da web: um número de pessoas geograficamente dispersas com interesses similares podem se comunicar
  - Comunidade dos pais de crianças com hemofilia
  - Pessoas interessadas em traduzir textos em Latim

Agregação de interesses

- Característica única da web: um número de pessoas geograficamente dispersas com interesses similares podem se comunicar
  - Comunidade dos pais de crianças com hemofilia
  - Pessoas interessadas em traduzir textos em Latim
  - Buscadores são incentivadores para agregação de interesses

ORI na web vs. ORI em geral

- Na web, busca não é simplesmente um característica boa

ORI na web vs. ORI em geral

- Na web, busca não é simplesmente um característica boa

ORI na web vs. ORI em geral

- Na web, busca não é simplesmente um característica boa
  - busca facilita para a web: ...

ORI na web vs. ORI em geral

- Na web, busca não é simplesmente um característica boa
  - busca facilita para a web: ...
  - ...financiamento, criação de conteúdo, agregação de interesses etc.

ORI na web vs. ORI em geral

- Na web, busca não é simplesmente um característica boa
  - busca facilita para a web: ...
  - ...financiamento, criação de conteúdo, agregação de interesses etc.
- Web é uma coleção caótica e descoordenada

ORI na web vs. ORI em geral

- Na web, busca não é simplesmente um característica boa
  - busca facilita para a web: ...
  - ...financiamento, criação de conteúdo, agregação de interesses etc.
- Web é uma coleção caótica e descoordenada
- Sem controle e restrições de quem pode criar conteúdo

ORI na web vs. ORI em geral

- Na web, busca não é simplesmente um característica boa
  - busca facilita para a web: ...
  - ...financiamento, criação de conteúdo, agregação de interesses etc.
- Web é uma coleção caótica e descoordenada
- Sem controle e restrições de quem pode criar conteúdo
- A web é muito extensa

ORI na web vs. ORI em geral

- Na web, busca não é simplesmente um característica boa
  - busca facilita para a web: ...
  - ...financiamento, criação de conteúdo, agregação de interesses etc.

  $\rightarrow$ ver anúncios
- Web é uma coleção caótica e descoordenada
- Sem controle e restrições de quem pode criar conteúdo
- A web é muito extensa

ORI na web vs. ORI em geral

- Na web, busca não é simplesmente um característica boa
  - busca facilita para a web: ...
  - ...financiamento, criação de conteúdo, agregação de interesses etc.

  → ver anúncios
- Web é uma coleção caótica e descoordenada → duplicatas – necessário detectar
- Sem controle e restrições de quem pode criar conteúdo

- A web é muito extensa

ORI na web vs. ORI em geral

- Na web, busca não é simplesmente um característica boa
  - busca facilita para a web: ...
  - ...financiamento, criação de conteúdo, agregação de interesses etc.

  → ver anúncios
- Web é uma coleção caótica e descoordenada → duplicatas – necessário detectar
- Sem controle e restrições de quem pode criar conteúdo → muito spam – necessário detectar spam
- A web é muito extensa

ORI na web vs. ORI em geral

- Na web, busca não é simplesmente um característica boa
  - busca facilita para a web: ...
  - ...financiamento, criação de conteúdo, agregação de interesses etc.

  → ver anúncios
- Web é uma coleção caótica e descoordenada → duplicatas – necessário detectar
- Sem controle e restrições de quem pode criar conteúdo → muito spam – necessário detectar spam
- A web é muito extensa → necessário saber o seu tamanho

Conteúdo

- De maneira geral

Conteúdo

- De maneira geral

Conteúdo

- De maneira geral
- Anúncios – pagam pela web

Conteúdo

- De maneira geral
- Anúncios – pagam pela web
- Detecção de duplicatas – problema da criação de conteúdo caótica

Conteúdo

- De maneira geral
- Anúncios – pagam pela web
- Detecção de duplicatas – problema da criação de conteúdo caótica
- Detecção de Spam – problema de falta de acessor central

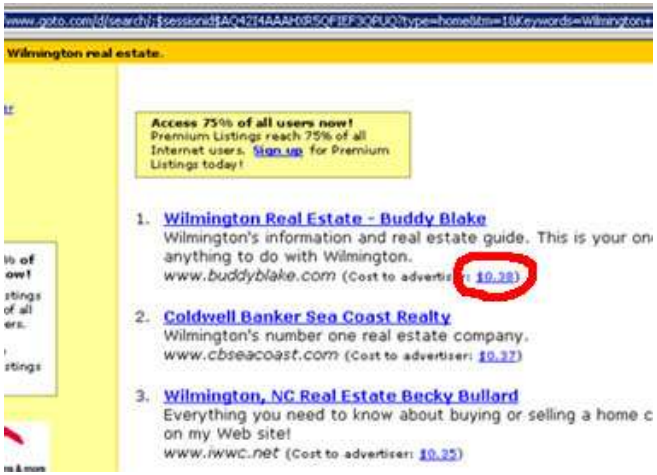
Conteúdo

- De maneira geral
- Anúncios – pagam pela web
- Detecção de duplicatas – problema da criação de conteúdo caótica
- Detecção de Spam – problema de falta de acessor central
  - Recuperação de informação da Web
  - Tamanho da web

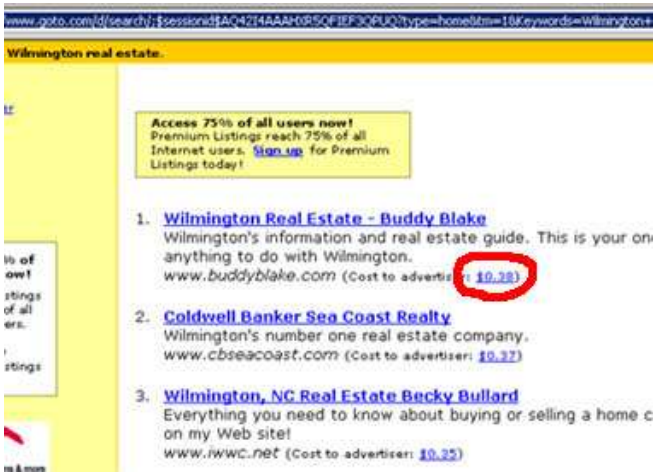
Conteúdo

Primeira geração de anúncios: Goto (1996)

Primeira geração de anúncios: Goto (1996)

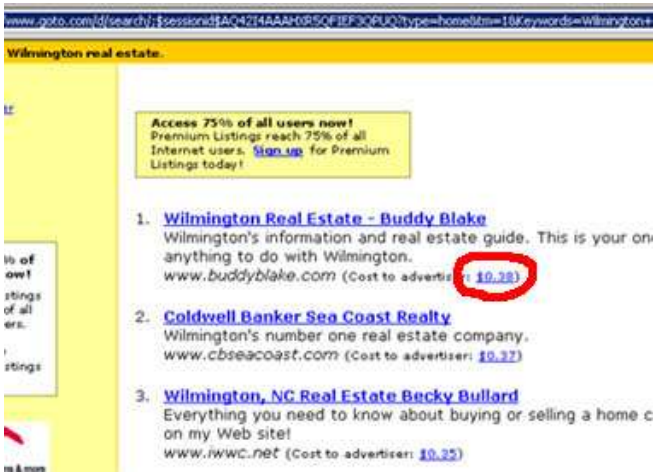


Primeira geração de anúncios: Goto (1996)



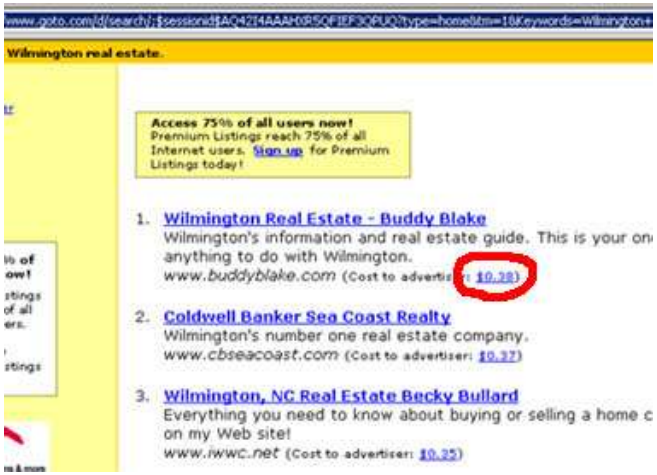
- Alguém ofertou o máximo ($0.38) por esta busca.

Primeira geração de anúncios: Goto (1996)



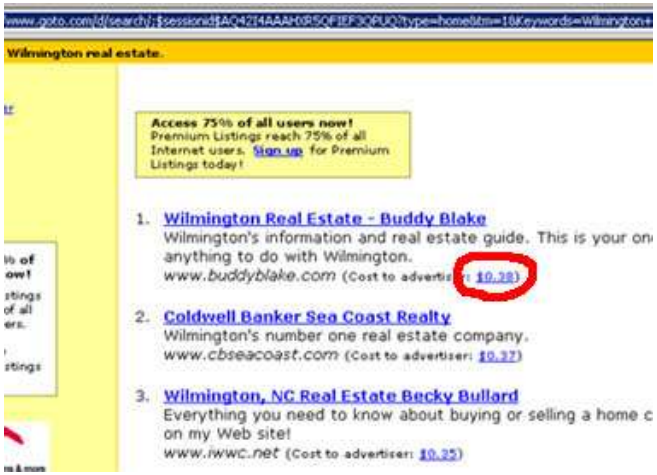
- Alguém ofertou o máximo ($0.38) por esta busca.
- Ele pagou $0.38 para o site Goto toda vez que alguém clicou no link.

Primeira geração de anúncios: Goto (1996)



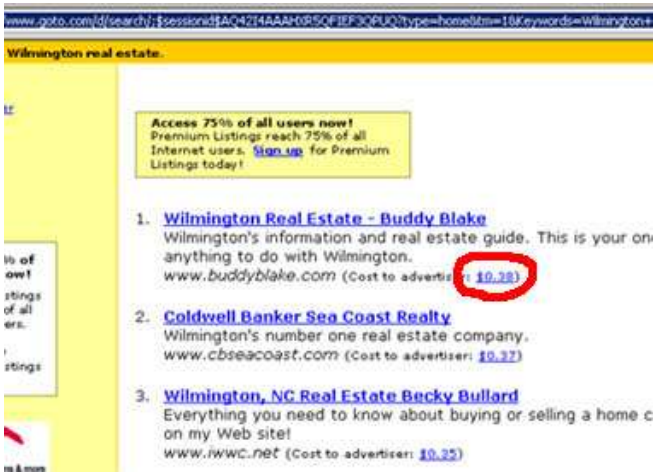
- Alguém ofertou o máximo ($0.38) por esta busca.
- Ele pagou $0.38 para o site Goto toda vez que alguém clicou no link.
- Página rankeada de acordo com ofertas – maximização de receita para o Goto

Primeira geração de anúncios: Goto (1996)



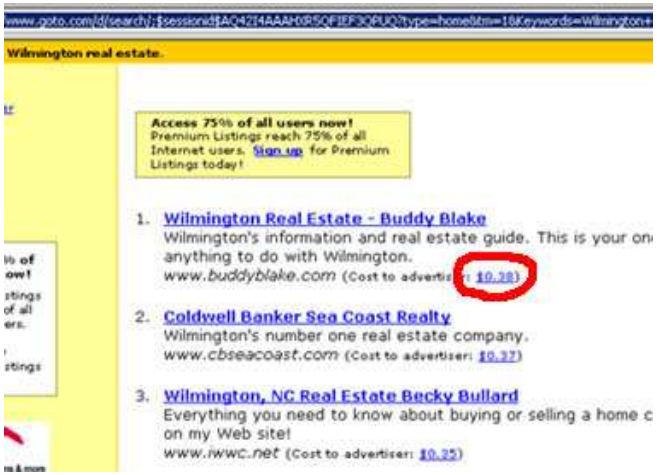
- Alguém ofertou o máximo ($0.38) por esta busca.
- Ele pagou $0.38 para o site Goto toda vez que alguém clicou no link.
- Página rankeada de acordo com ofertas – maximização de receita para o Goto
- Sem separação de anúncios/docs.

Primeira geração de anúncios: Goto (1996)



- Alguém ofertou o máximo ($0.38) por esta busca.
- Ele pagou $0.38 para o site Goto toda vez que alguém clicou no link.
- Página rankeada de acordo com ofertas – maximização de receita para o Goto
- Sem separação de anúncios/docs.
- Direto e honesto. Sem ranking de relevância, . . .

Primeira geração de anúncios: Goto (1996)

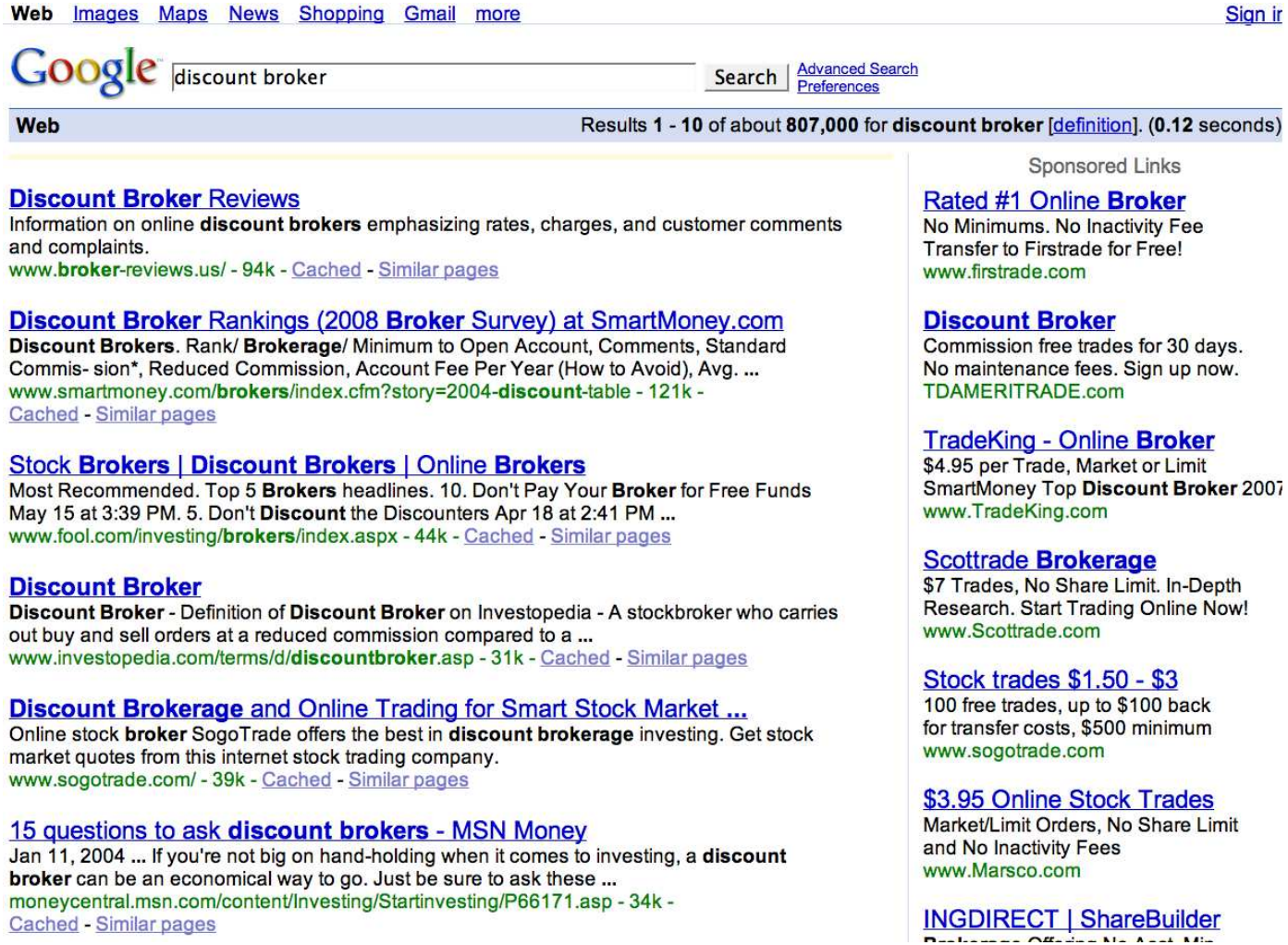


- Alguém ofertou o máximo ($0.38) por esta busca.
- Ele pagou $0.38 para o site Goto toda vez que alguém clicou no link.
- Página rankeada de acordo com ofertas – maximização de receita para o Goto
- Sem separação de anúncios/docs.
- Direto e honesto. Sem ranking de relevância, . . .
- . . . mas Goto não fazia como se houvesse tal ranking.

Segunda geração de anúncios: Google (2000/2001)

- Separação de resultados e anúncios

Dois rankings: páginas web (esq.) e anúncios (dir.)



Web Images Maps News Shopping Gmail more Sign in

Google discount broker Search Advanced Search Preferences

Web Results 1 - 10 of about 807,000 for discount broker [definition]. (0.12 seconds)

Discount Broker Reviews
Information on online **discount brokers** emphasizing rates, charges, and customer comments and complaints.
www.**broker**-reviews.us/ - 94k - Cached - Similar pages

Discount Broker Rankings (2008 Broker Survey) at SmartMoney.com
Discount Brokers. Rank/ **Brokerage**/ Minimum to Open Account, Comments, Standard Commis- sion*, Reduced Commission, Account Fee Per Year (How to Avoid), Avg. ...
www.smartmoney.com/**brokers**/index.cfm?story=2004-**discount**-table - 121k - Cached - Similar pages

Stock Brokers | Discount Brokers | Online Brokers
Most Recommended. Top 5 **Brokers** headlines. 10. Don't Pay Your **Broker** for Free Funds May 15 at 3:39 PM. 5. Don't **Discount** the Discounters Apr 18 at 2:41 PM ...
www.fool.com/investing/**brokers**/index.aspx - 44k - Cached - Similar pages

Discount Broker
Discount Broker - Definition of **Discount Broker** on Investopedia - A stockbroker who carries out buy and sell orders at a reduced commission compared to a ...
www.investopedia.com/terms/d/**discountbroker**.asp - 31k - Cached - Similar pages

Discount Brokerage and Online Trading for Smart Stock Market ...
Online stock **broker** SogoTrade offers the best in **discount brokerage** investing. Get stock market quotes from this internet stock trading company.
www.sogotrade.com/ - 39k - Cached - Similar pages

15 questions to ask discount brokers - MSN Money
Jan 11, 2004 ... If you're not big on hand-holding when it comes to investing, a **discount broker** can be an economical way to go. Just be sure to ask these ...
moneycentral.msn.com/content/Investing/Startinvesting/P66171.asp - 34k - Cached - Similar pages

Sponsored Links

Rated #1 Online Broker
No Minimums. No Inactivity Fee
Transfer to Firstrade for Free!
www.firstrade.com

Discount Broker
Commission free trades for 30 days.
No maintenance fees. Sign up now.
TDAMERITRADE.com

TradeKing - Online Broker
$4.95 per Trade, Market or Limit
SmartMoney Top **Discount Broker** 2007
www.TradeKing.com

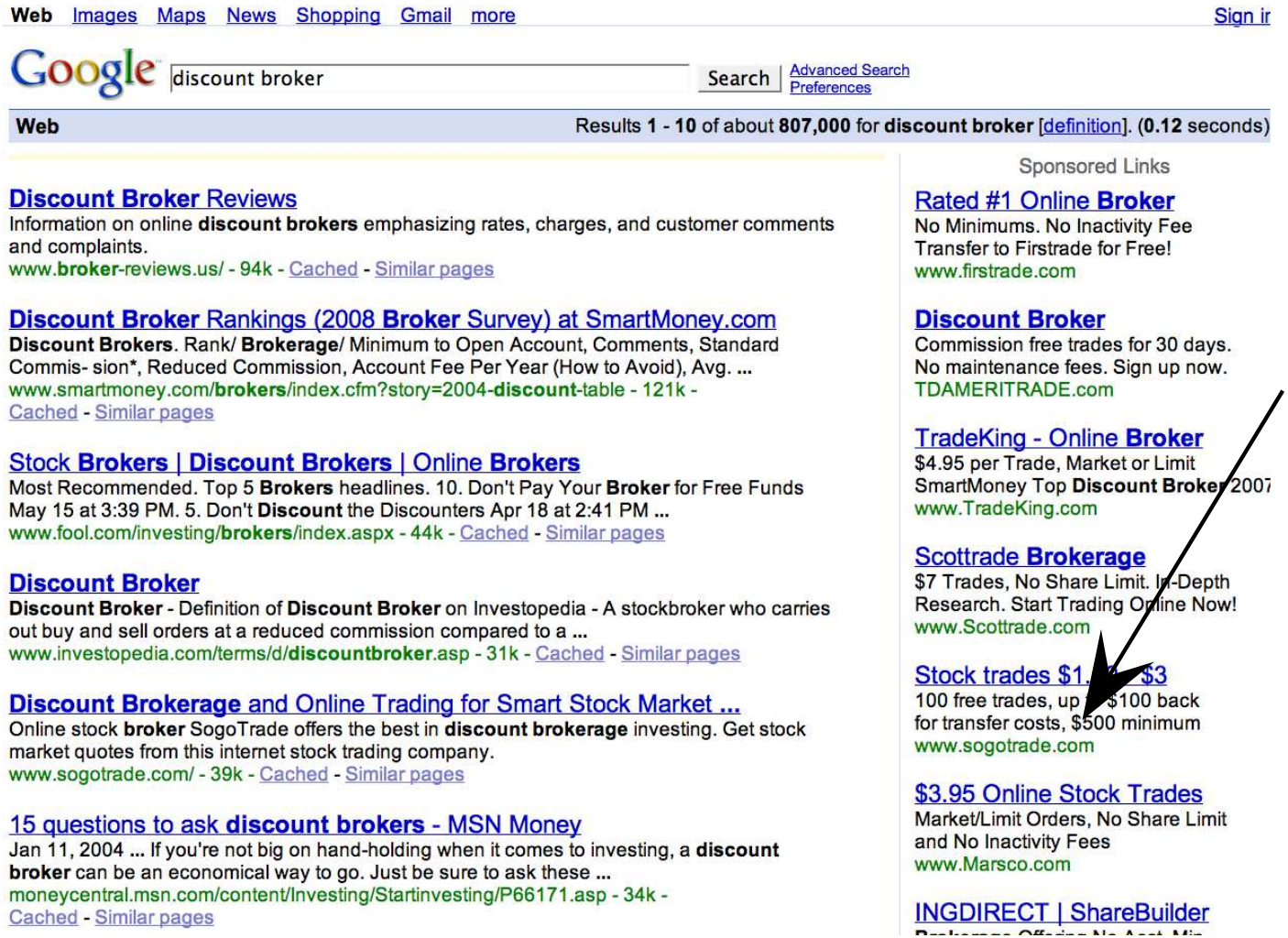
Scottrade Brokerage
$7 Trades, No Share Limit. In-Depth Research. Start Trading Online Now!
www.Scottrade.com

Stock trades $1.50 - $3
100 free trades, up to $100 back for transfer costs, $500 minimum
www.sogotrade.com

$3.95 Online Stock Trades
Market/Limit Orders, No Share Limit and No Inactivity Fees
www.Marsco.com

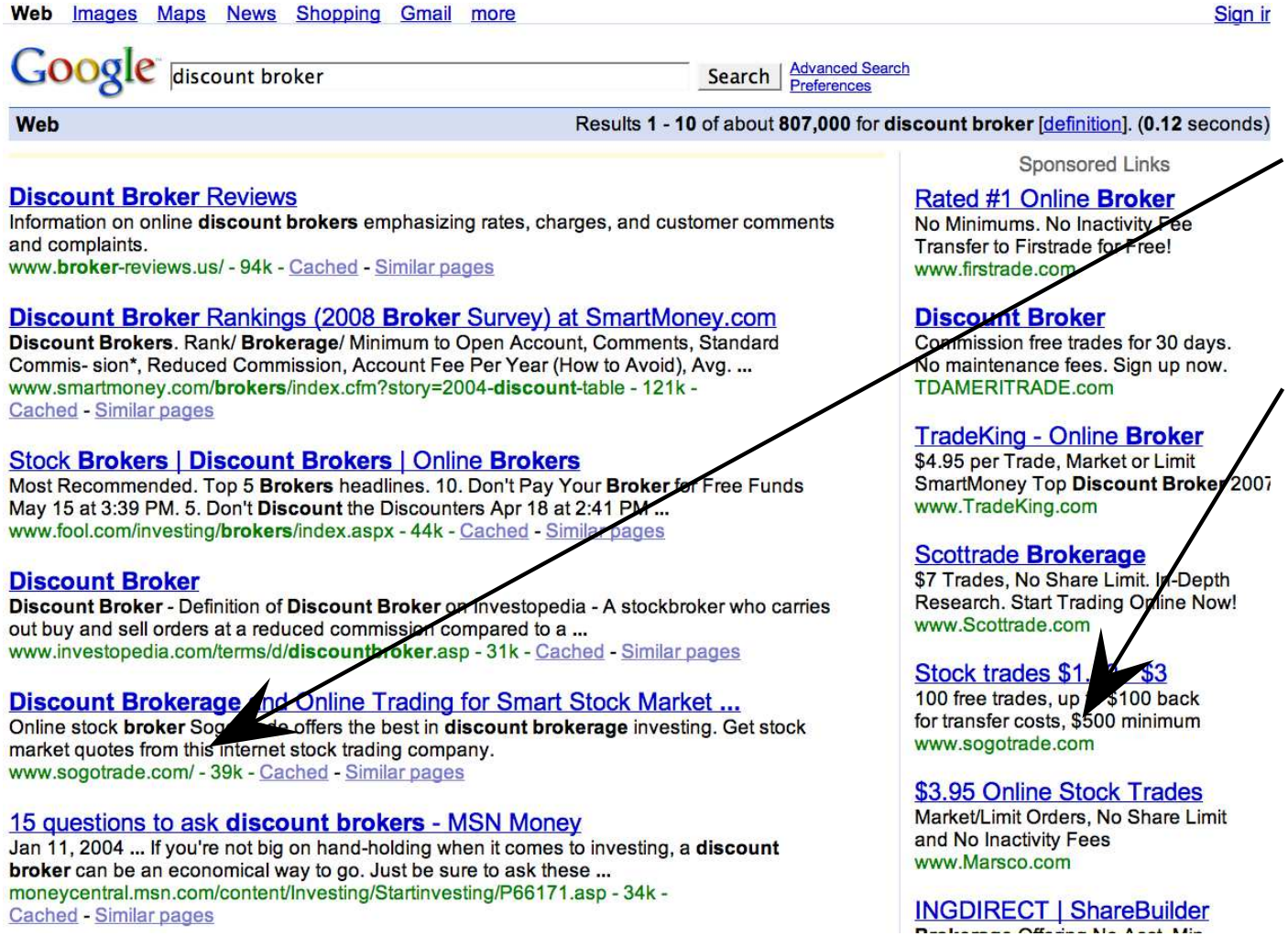
INGDIRECT | ShareBuilder

Dois rankings: páginas web (esq.) e anúncios (dir.)



SogoTrade aparece nos anúncios.

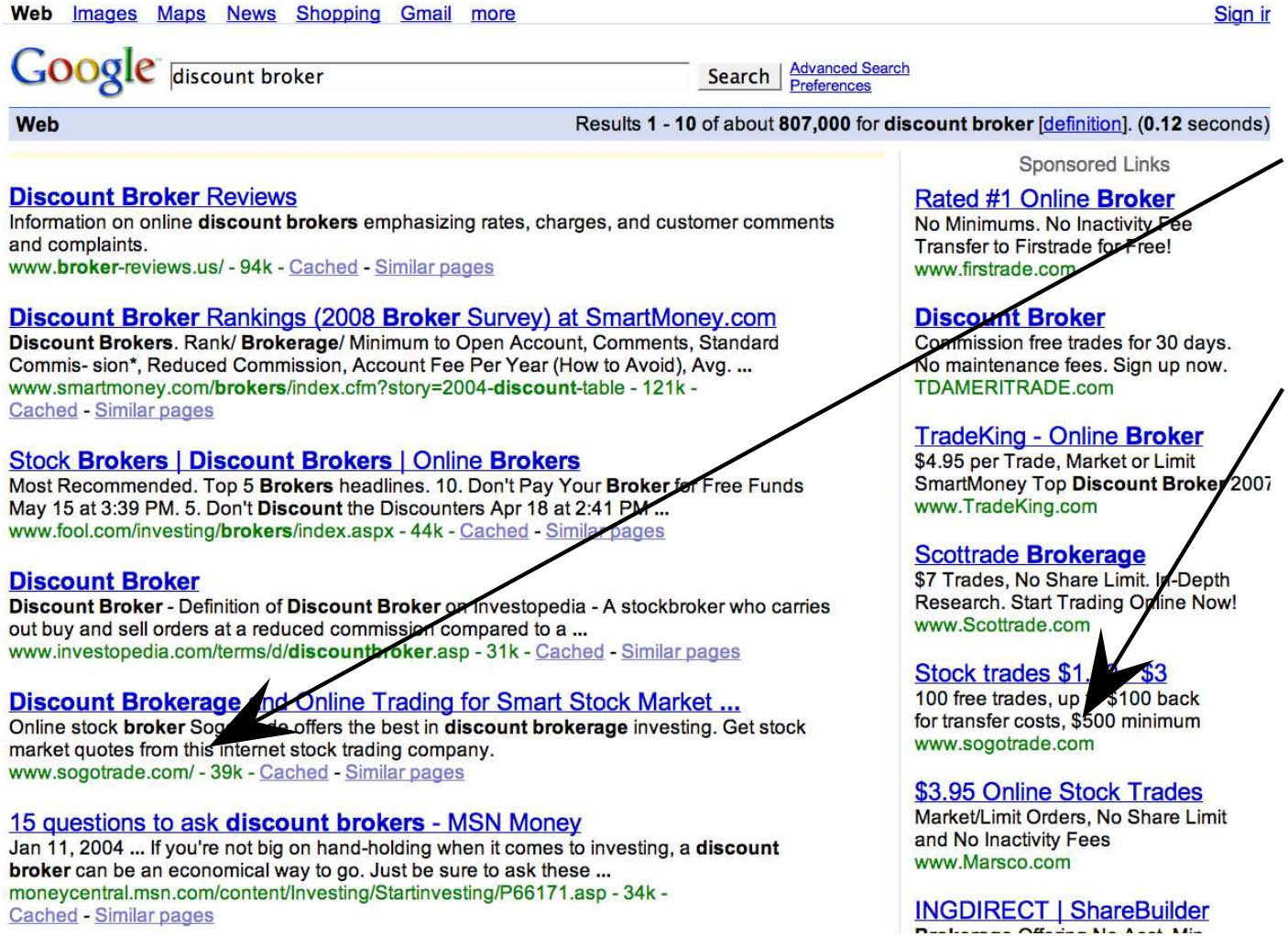
Dois rankings: páginas web (esq.) e anúncios (dir.)



SogoTrade aparece nos resultados de busca.

SogoTrade aparece nos anúncios.

Dois rankings: páginas web (esq.) e anúncios (dir.)

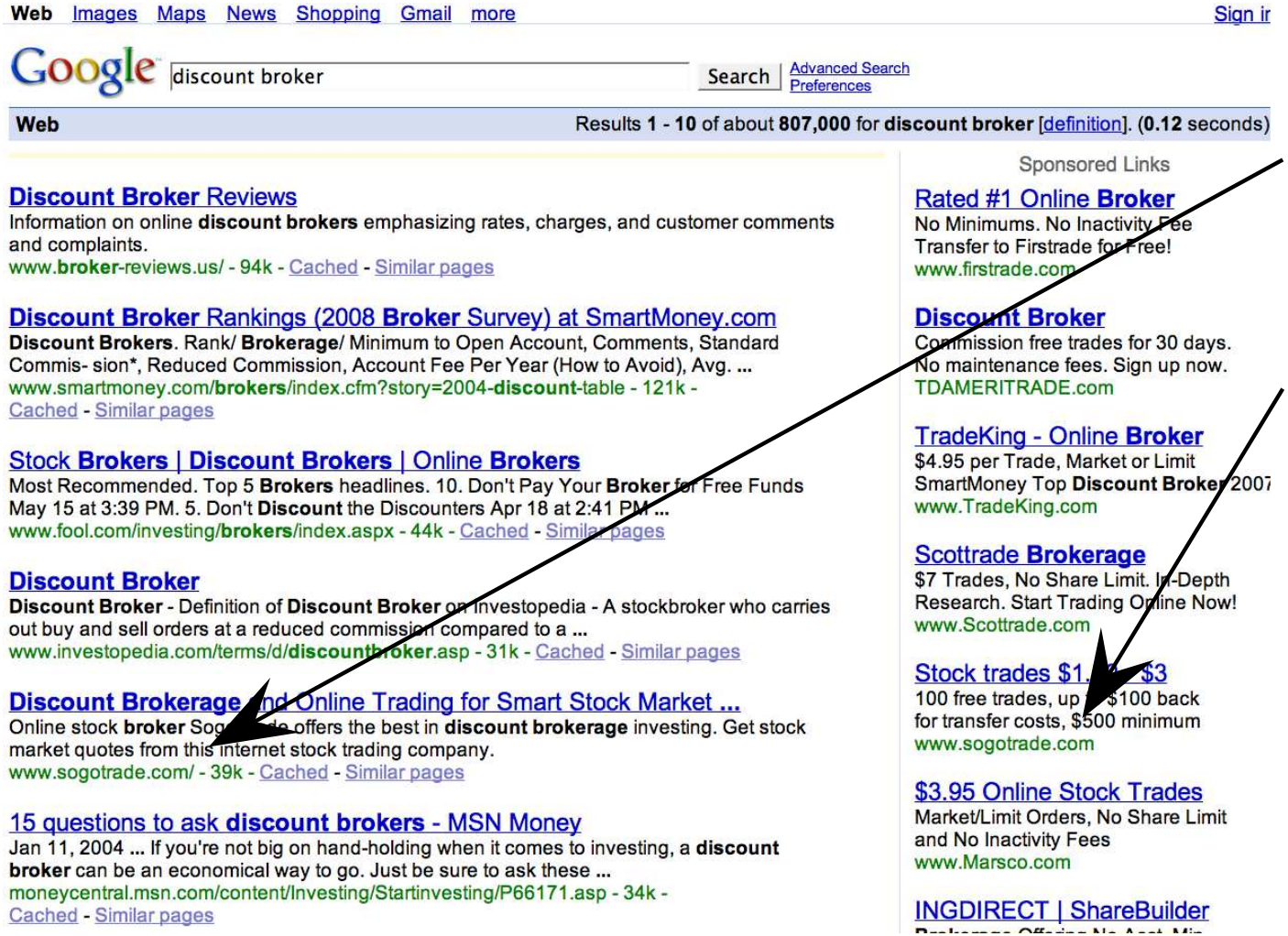


SogoTrade aparece nos resultados de busca.

SogoTrade aparece nos anúncios.

buscadores posicionam melhor no ranking anuncionantes que não anuncionantes?

Dois rankings: páginas web (esq.) e anúncios (dir.)



SogoTrade aparece nos resultados de busca.

SogoTrade aparece nos anúncios.

buscadores posicionam melhor no ranking anuncionantes que não anuncionantes?

Buscadores dizem que não.

Anúncios influenciam o conteúdo editorial?

- Problema similar em jornais e canais de TV

Anúncios influenciam o conteúdo editorial?

- Problema similar em jornais e canais de TV
- Um jornal é relutante em publicar críticas duras aos seus anunciantes

Como anúncios são rankeados?

- Anunciantes fazem ofertas/lances por palavras-chaves – venda por leilão.

	Buscas no Google (comScore Inc.)	Média diária
2013	2,161,530,000,000	5,922,000,000
2012	1,873,910,000,000	5,134,000,000
2011	1,722,071,000,000	4,717,000,000

Como anúncios são rankeados?

- Anunciantes fazem ofertas/lances por palavras-chaves – venda por leilão.
- Sistema aberto: qualquer um pode participar e fazer lances em palavras-chaves

	Buscas no Google (comScore Inc.)	Média diária
2013	2,161,530,000,000	5,922,000,000
2012	1,873,910,000,000	5,134,000,000
2011	1,722,071,000,000	4,717,000,000

Como anúncios são rankeados?

- Anunciantes fazem ofertas/lances por palavras-chaves – venda por leilão.
- Sistema aberto: qualquer um pode participar e fazer lances em palavras-chaves
- Anunciantes são cobrados somente quando alguém clica no anúncio

	Buscas no Google (comScore Inc.)	Média diária
2013	2,161,530,000,000	5,922,000,000
2012	1,873,910,000,000	5,134,000,000
2011	1,722,071,000,000	4,717,000,000

Como anúncios são rankeados?

- Anunciantes fazem ofertas/lances por palavras-chaves – venda por leilão.
- Sistema aberto: qualquer um pode participar e fazer lances em palavras-chaves
- Anunciantes são cobrados somente quando alguém clica no anúncio
- Como o leilão determina o ranking de um anúncio e o preço pago pelo anúncio?

	Buscas no Google (comScore Inc.)	Média diária
2013	2,161,530,000,000	5,922,000,000
2012	1,873,910,000,000	5,134,000,000
2011	1,722,071,000,000	4,717,000,000

Como anúncios são rankeados?

- Anunciantes fazem ofertas/lances por palavras-chaves – venda por leilão.
- Sistema aberto: qualquer um pode participar e fazer lances em palavras-chaves
- Anunciantes são cobrados somente quando alguém clica no anúncio
- Como o leilão determina o ranking de um anúncio e o preço pago pelo anúncio?
- A base é um leilão de segundo preço, mas com alterações

	Buscas no Google (comScore Inc.)	Média diária
2013	2,161,530,000,000	5,922,000,000
2012	1,873,910,000,000	5,134,000,000
2011	1,722,071,000,000	4,717,000,000

Como anúncios são rankeados?

- Anunciantes fazem ofertas/lances por palavras-chaves – venda por leilão.
- Sistema aberto: qualquer um pode participar e fazer lances em palavras-chaves
- Anunciantes são cobrados somente quando alguém clica no anúncio
- Como o leilão determina o ranking de um anúncio e o preço pago pelo anúncio?
- A base é um leilão de segundo preço, mas com alterações
- Tópico importante de pesquisa – publicidade computacional

	Buscas no Google (comScore Inc.)	Média diária
2013	2,161,530,000,000	5,922,000,000
2012	1,873,910,000,000	5,134,000,000
2011	1,722,071,000,000	4,717,000,000

Como anúncios são rankeados?

- Anunciantes fazem ofertas/lances por palavras-chaves – venda por leilão.
- Sistema aberto: qualquer um pode participar e fazer lances em palavras-chaves
- Anunciantes são cobrados somente quando alguém clica no anúncio
- Como o leilão determina o ranking de um anúncio e o preço pago pelo anúncio?
- A base é um leilão de segundo preço, mas com alterações
- Tópico importante de pesquisa – publicidade computacional
  - Aumentar um centavo a mais de cada anúncio significa bilhões de receita adicional para o buscador.

	Buscas no Google (comScore Inc.)	Média diária
2013	2,161,530,000,000	5,922,000,000
2012	1,873,910,000,000	5,134,000,000
2011	1,722,071,000,000	4,717,000,000

Como anúncios são rankeados?

- Primeira parte: de acordo com valores do lance à la Goto

Como anúncios são rankeados?

- Primeira parte: de acordo com valores do lance à la Goto
  - Má ideia: suscetível a manipulação

Como anúncios são rankeados?

- Primeira parte: de acordo com valores do lance à la Goto
  - Má ideia: suscetível a manipulação
  - Exemplo: consulta [cancer] $\rightarrow$ anúncio de funerária

Como anúncios são rankeados?

- Primeira parte: de acordo com valores do lance à la Goto
  - Má ideia: suscetível a manipulação
  - Exemplo: consulta [cancer] $\rightarrow$ anúncio de funerária
  - Não queremos mostrar anúncios não-relevantes ou ofensivos

Como anúncios são rankeados?

- Primeira parte: de acordo com valores do lance à la Goto
  - Má ideia: suscetível a manipulação
  - Exemplo: consulta [cancer] $\rightarrow$ anúncio de funerária
  - Não queremos mostrar anúncios não-relevantes ou ofensivos
- Em vez disso: ordenar baseado nos valores dos lances e uma medida de qualidade

Como anúncios são rankeados?

- Primeira parte: de acordo com valores do lance à la Goto
  - Má ideia: suscetível a manipulação
  - Exemplo: consulta [cancer] $\rightarrow$ anúncio de funerária
  - Não queremos mostrar anúncios não-relevantes ou ofensivos
- Em vez disso: ordenar baseado nos valores dos lances e uma medida de qualidade
- Qualidade = relevância + página de chegada + taxa de cliques

Como anúncios são rankeados?

- Primeira parte: de acordo com valores do lance à la Goto
  - Má ideia: suscetível a manipulação
  - Exemplo: consulta [cancer] $\rightarrow$ anúncio de funerária
  - Não queremos mostrar anúncios não-relevantes ou ofensivos
- Em vez disso: ordenar baseado nos valores dos lances e uma medida de qualidade
- Qualidade = relevância + página de chegada + taxa de cliques
  - taxa de cliques = TDC = números de cliques por número total de exibição

Como anúncios são rankeados?

- Primeira parte: de acordo com valores do lance à la Goto
  - Má ideia: suscetível a manipulação
  - Exemplo: consulta [cancer] $\rightarrow$ anúncio de funerária
  - Não queremos mostrar anúncios não-relevantes ou ofensivos
- Em vez disso: ordenar baseado nos valores dos lances e uma medida de qualidade
- Qualidade = relevância + página de chegada + taxa de cliques
  - taxa de cliques = TDC = números de cliques por número total de exibição
- Resultado: um anúncio não relevante será rankeado baixo

Como anúncios são rankeados?

- Primeira parte: de acordo com valores do lance à la Goto
  - Má ideia: suscetível a manipulação
  - Exemplo: consulta [cancer] $\rightarrow$ anúncio de funerária
  - Não queremos mostrar anúncios não-relevantes ou ofensivos
- Em vez disso: ordenar baseado nos valores dos lances e uma medida de qualidade
- Qualidade = relevância + página de chegada + taxa de cliques
  - taxa de cliques = TDC = números de cliques por número total de exibição
- Resultado: um anúncio não relevante será rankeado baixo
  - Mesmo se isso diminuir a receita a curto-prazo

Como anúncios são rankeados?

- Primeira parte: de acordo com valores do lance à la Goto
  - Má ideia: suscetível a manipulação
  - Exemplo: consulta [cancer] $\rightarrow$ anúncio de funerária
  - Não queremos mostrar anúncios não-relevantes ou ofensivos
- Em vez disso: ordenar baseado nos valores dos lances e uma medida de qualidade
- Qualidade = relevância + página de chegada + taxa de cliques
  - taxa de cliques = TDC = números de cliques por número total de exibição
- Resultado: um anúncio não relevante será rankeado baixo
  - Mesmo se isso diminuir a receita a curto-prazo
  - Esperança: aceitação geral do sistema e receita geral é maximizada se usuários recebem informação útil

Como anúncios são rankeados?

- Primeira parte: de acordo com valores do lance à la Goto
  - Má ideia: suscetível a manipulação
  - Exemplo: consulta [cancer] $\rightarrow$ anúncio de funerária
  - Não queremos mostrar anúncios não-relevantes ou ofensivos
- Em vez disso: ordenar baseado nos valores dos lances e uma medida de qualidade
- Qualidade = relevância + página de chegada + taxa de cliques
  - taxa de cliques = TDC = números de cliques por número total de exibição
- Resultado: um anúncio não relevante será rankeado baixo
  - Mesmo se isso diminuir a receita a curto-prazo
  - Esperança: aceitação geral do sistema e receita geral é maximizada se usuários recebem informação útil
- Outros fatores de ranking: localidades, horas do dia, qualidade e tempo de carregamento da página

Como anúncios são rankeados?

- Primeira parte: de acordo com valores do lance à la Goto
  - Má ideia: suscetível a manipulação
  - Exemplo: consulta [cancer] $\rightarrow$ anúncio de funerária
  - Não queremos mostrar anúncios não-relevantes ou ofensivos
- Em vez disso: ordenar baseado nos valores dos lances e uma medida de qualidade
- Qualidade = relevância + página de chegada + taxa de cliques
  - taxa de cliques = TDC = números de cliques por número total de exibição
- Resultado: um anúncio não relevante será rankeado baixo
  - Mesmo se isso diminuir a receita a curto-prazo
  - Esperança: aceitação geral do sistema e receita geral é maximizada se usuários recebem informação útil
- Outros fatores de ranking: localidades, horas do dia, qualidade e tempo de carregamento da página
- O fator principal de ranking: a consulta

Leilão de segunda posição do Google (simplificado)

anunciante	lance	TDC	valor rank	rank	preço
A	$4.00	0.01	0.04	4	(mínimo)
B	$3.00	0.03	0.09	2	$2.68
C	$2.00	0.06	0.12	1	$1.51
D	$1.00	0.08	0.08	3	$0.51

Leilão de segunda posição do Google (simplificado)

anunciante	lance	TDC	valor rank	rank	preço
A	\$4.00	0.01	0.04	4	(mínimo)
B	\$3.00	0.03	0.09	2	\$2.68
C	\$2.00	0.06	0.12	1	\$1.51
D	\$1.00	0.08	0.08	3	\$0.51

- lance: lance máximo de um clique pelo anunciante
- TDC: taxa de cliques: quando um anúncio é mostrado, qual porcentagem de vezes que usuários clicam nele? TDC é uma medida de relevância.
- rank: rank no leilão
- preço: pago pelo anunciante

$$\text{preço}_{\text{rank}} = \text{lance}_{\text{rank}+1} \frac{\text{TDC}_{\text{rank}+1}}{\text{TDC}_{\text{rank}}}$$

Leilão de segunda posição do Google (simplificado)

anunciante	lance	TDC	valor rank	rank	preço
A	\$4.00	0.01	0.04	4	(mínimo)
B	\$3.00	0.03	0.09	2	\$2.68
C	\$2.00	0.06	0.12	1	\$1.51
D	\$1.00	0.08	0.08	3	\$0.51

Leilão de segundo preço: O anunciante paga a menor quantidade necessária para manter a posição no leilão (mais 1 centavo).

$\text{preço}_1 \times \text{TDC}_1 = \text{lance}_2 \times \text{TDC}_2$ (resulta em $\text{rank}_1 = \text{rank}_2$)

$\text{preço}_1 = \text{lance}_2 \times \text{TDC}_2 \,/\, \text{TDC}_1$

$p_1 = \text{lance}_2 \times \text{TDC}_2/\text{TDC}_1 = 3.00 \times 0.03/0.06 = 1.50$
$p_2 = \text{lance}_3 \times \text{TDC}_3/\text{TDC}_2 = 1.00 \times 0.08/0.03 = 2.67$
$p_3 = \text{lance}_4 \times \text{TDC}_4/\text{TDC}_3 = 4.00 \times 0.01/0.08 = 0.50$

Anúncios

- O buscador recebe receita toda vez que alguém clica no anúncio

Anúncios

- O buscador recebe receita toda vez que alguém clica no anúncio
- O usuário somente clica em um anúncio se está interessado nele

Anúncios

- O buscador recebe receita toda vez que alguém clica no anúncio
- O usuário somente clica em um anúncio se está interessado nele
  - Buscadores punem anúncios enganosos e não relevantes

Anúncios

- O buscador recebe receita toda vez que alguém clica no anúncio
- O usuário somente clica em um anúncio se está interessado nele
  - Buscadores punem anúncios enganosos e não relevantes
  - Resultado: usuários frequentemente estão satisfeitos após clicar em um anúncio

Anúncios

- O buscador recebe receita toda vez que alguém clica no anúncio
- O usuário somente clica em um anúncio se está interessado nele
  - Buscadores punem anúncios enganosos e não relevantes
  - Resultado: usuários frequentemente estão satisfeitos após clicar em um anúncio
- O anunciante encontra novos clientes a um bom preço

Nem todos ganham: negociação de palavras-chaves

- Alguém compra uma palavra-chave no Google

Nem todos ganham: negociação de palavras-chaves

- Alguém compra uma palavra-chave no Google
- E redireciona tráfego para outra pessoa, pagando muito mais que pagou ao Google

Nem todos ganham: negociação de palavras-chaves

- Alguém compra uma palavra-chave no Google
- E redireciona tráfego para outra pessoa, pagando muito mais que pagou ao Google
  - E.g., redirecionar para uma página cheia de anúncios

Nem todos ganham: negociação de palavras-chaves

- Alguém compra uma palavra-chave no Google
- E redireciona tráfego para outra pessoa, pagando muito mais que pagou ao Google
  - E.g., redirecionar para uma página cheia de anúncios
- Isto atrapalha usuários

Nem todos ganham: negociação de palavras-chaves

- Alguém compra uma palavra-chave no Google
- E redireciona tráfego para outra pessoa, pagando muito mais que pagou ao Google
  - E.g., redirecionar para uma página cheia de anúncios
- Isto atrapalha usuários
- Spammers sempre inventam novas formas

Nem todos ganham: negociação de palavras-chaves

- Alguém compra uma palavra-chave no Google
- E redireciona tráfego para outra pessoa, pagando muito mais que pagou ao Google
  - E.g., redirecionar para uma página cheia de anúncios
- Isto atrapalha usuários
- Spammers sempre inventam novas formas
- Buscadores sempre tentam combater essas novas formas

Nem todos ganham: Violação de marcas

- Exemplo: geico

Nem todos ganham: Violação de marcas

- Exemplo: geico
- Durante 2005: o termo "geico" no Google foi comprado por competidores

Nem todos ganham: Violação de marcas

- Exemplo: geico
- Durante 2005: o termo "geico" no Google foi comprado por competidores
- Geico perdeu caso nos EUA

Nem todos ganham: Violação de marcas

- Exemplo: geico
- Durante 2005: o termo "geico" no Google foi comprado por competidores
- Geico perdeu caso nos EUA
- Louis Vuitton perdeu caso parecido na Europa.

Nem todos ganham: Violação de marcas

- Exemplo: geico
- Durante 2005: o termo "geico" no Google foi comprado por competidores
- Geico perdeu caso nos EUA
- Louis Vuitton perdeu caso parecido na Europa.
- Ver http://google.com/tm_complaint.html

Conteúdo

Detecção de duplicatas

- A web está cheia de conteúdo duplicado

Detecção de duplicatas

- A web está cheia de conteúdo duplicado
- Mais que muitas outras coleções

Detecção de duplicatas

- A web está cheia de conteúdo duplicado
- Mais que muitas outras coleções
- Duplicatas exatas

Detecção de duplicatas

- A web está cheia de conteúdo duplicado
- Mais que muitas outras coleções
- Duplicatas exatas
  - Fácil de eliminar

Detecção de duplicatas

- A web está cheia de conteúdo duplicado
- Mais que muitas outras coleções
- Duplicatas exatas
  - Fácil de eliminar
  - E.g., usar hash/"impressão digital"

Detecção de duplicatas

- A web está cheia de conteúdo duplicado
- Mais que muitas outras coleções
- Duplicatas exatas
  - Fácil de eliminar
  - E.g., usar hash/"impressão digital"
- Quase-duplicatas são abundantes na web

Detecção de duplicatas

- A web está cheia de conteúdo duplicado
- Mais que muitas outras coleções
- Duplicatas exatas
  - Fácil de eliminar
  - E.g., usar hash/"impressão digital"
- Quase-duplicatas são abundantes na web
- Para usuário, uma busca com documentos quase idênticos é ruim

Detecção de duplicatas

- A web está cheia de conteúdo duplicado
- Mais que muitas outras coleções
- Duplicatas exatas
  - Fácil de eliminar
  - E.g., usar hash/"impressão digital"
- Quase-duplicatas são abundantes na web
- Para usuário, uma busca com documentos quase idênticos é ruim
- Quase-duplicatas não são relevantes e devem ser eliminadas

Detecção de duplicatas

- A web está cheia de conteúdo duplicado
- Mais que muitas outras coleções
- Duplicatas exatas
  - Fácil de eliminar
  - E.g., usar hash/"impressão digital"
- Quase-duplicatas são abundantes na web
- Para usuário, uma busca com documentos quase idênticos é ruim
- Quase-duplicatas não são relevantes e devem ser eliminadas
  - Difíceis de eliminar

Exemplo: quase-duplicatas

Detecção quase-duplicatas

- Calcular similaridade com distância de edição

Detecção quase-duplicatas

- Calcular similaridade com distância de edição
- Queremos similaridade "sintática" (em vez de semântica).

Detecção quase-duplicatas

- Calcular similaridade com distância de edição
- Queremos similaridade "sintática" (em vez de semântica).
  - Similaridade semântica (conteúdo) é muito difícil calcular

Detecção quase-duplicatas

- Calcular similaridade com distância de edição
- Queremos similaridade "sintática" (em vez de semântica).
  - Similaridade semântica (conteúdo) é muito difícil calcular
- Não consideramos documentos como quase-duplicatas se tem o mesmo conteúdo, mas se expressando com diferentes palavras

Detecção quase-duplicatas

- Calcular similaridade com distância de edição
- Queremos similaridade "sintática" (em vez de semântica).
  - Similaridade semântica (conteúdo) é muito difícil calcular
- Não consideramos documentos como quase-duplicatas se tem o mesmo conteúdo, mas se expressando com diferentes palavras
- Usar limiar de similaridade θ para decidir se "é/não é uma quase duplicata"

Detecção quase-duplicatas

- Calcular similaridade com distância de edição
- Queremos similaridade "sintática" (em vez de semântica).
  - Similaridade semântica (conteúdo) é muito difícil calcular
- Não consideramos documentos como quase-duplicatas se tem o mesmo conteúdo, mas se expressando com diferentes palavras
- Usar limiar de similaridade θ para decidir se "é/não é uma quase duplicata"
- E.g., dois docs são quase-duplicatas se similaridade $> \theta = 80\%$.

Representar cada documento como um conjunto de "composições"

- Uma composição é uma palavra n-grama.

Representar cada documento como um conjunto de "composições"

- Uma composição é uma palavra n-grama.
- Composições são usadas para medir similaridade sintática de documentos.

Representar cada documento como um conjunto de "composições"

- Uma composição é uma palavra n-grama.
- Composições são usadas para medir similaridade sintática de documentos.
- Por exemplo, para $n = 3$, "Rosa é a rosa é a rosa" pode ser representado como um conjunto de composições:

Representar cada documento como um conjunto de "composições"

- Uma composição é uma palavra n-grama.
- Composições são usadas para medir similaridade sintática de documentos.
- Por exemplo, para $n = 3$, "Rosa é a rosa é a rosa" pode ser representado como um conjunto de composições:
  - { a-rosa-é, rosa-é-a, é-a-rosa }

Representar cada documento como um conjunto de "composições"

- Uma composição é uma palavra n-grama.
- Composições são usadas para medir similaridade sintática de documentos.
- Por exemplo, para $n = 3$, "Rosa é a rosa é a rosa" pode ser representado como um conjunto de composições:
  - { a-rosa-é, rosa-é-a, é-a-rosa }
- Definimos a similaridade de dois documentos como o coeficiente de Jaccard dos respectivos conjuntos de composições.

Lembrando: coeficiente de Jaccard

- Um medida para sobreposição de dois conjuntos

Lembrando: coeficiente de Jaccard

- Um medida para sobreposição de dois conjuntos
- Sejam os dois conjuntos A e B

Lembrando: coeficiente de Jaccard

- Um medida para sobreposição de dois conjuntos
- Sejam os dois conjuntos A e B
- Coeficiente de Jaccard:

$$\textsc{jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

$(A \neq \emptyset$ ou $B \neq \emptyset)$

Lembrando: coeficiente de Jaccard

- Um medida para sobreposição de dois conjuntos
- Sejam os dois conjuntos A e B
- Coeficiente de Jaccard:

$$\textsc{jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

  $(A \neq \emptyset$ ou $B \neq \emptyset)$
- $\textsc{jaccard}(A, A) = 1$

Lembrando: coeficiente de Jaccard

- Um medida para sobreposição de dois conjuntos
- Sejam os dois conjuntos A e B
- Coeficiente de Jaccard:

$$\textsc{jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

  $(A \neq \emptyset$ ou $B \neq \emptyset)$
- $\textsc{jaccard}(A, A) = 1$
- $\textsc{jaccard}(A, B) = 0$ if $A \cap B = 0$

Lembrando: coeficiente de Jaccard

- Um medida para sobreposição de dois conjuntos
- Sejam os dois conjuntos A e B
- Coeficiente de Jaccard:

$$\textsc{jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

  $(A \neq \emptyset$ ou $B \neq \emptyset)$
- $\textsc{jaccard}(A, A) = 1$
- $\textsc{jaccard}(A, B) = 0$ if $A \cap B = 0$
- A e B não necessitam ter o mesmo tamanho

Lembrando: coeficiente de Jaccard

- Um medida para sobreposição de dois conjuntos
- Sejam os dois conjuntos A e B
- Coeficiente de Jaccard:

$$\textsc{jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

  $(A \neq \emptyset$ ou $B \neq \emptyset)$
- $\textsc{jaccard}(A, A) = 1$
- $\textsc{jaccard}(A, B) = 0$ if $A \cap B = 0$
- A e B não necessitam ter o mesmo tamanho
- Valores entre 0 e 1.

Exemplo: coeficiente de Jaccard

- Três documentos:
  d_1: "Jack London viajou para Oakland"
  d_2: "Jack London viajou para a cidade de Oakland"
  d_3: "Jack viajou de Oakland para London"

Exemplo: coeficiente de Jaccard

- Três documentos:
  d_1: "Jack London viajou para Oakland"
  d_2: "Jack London viajou para a cidade de Oakland"
  d_3: "Jack viajou de Oakland para London"
- Usando composições de tamanho 2 (2-gramas ou bigramas), quais são os coeficientes de Jaccard $J(d_1, d_2)$ e $J(d_1, d_3)$?

Exemplo: coeficiente de Jaccard

- Três documentos:
  d_1: "Jack London viajou para Oakland"
  d_2: "Jack London viajou para a cidade de Oakland"
  d_3: "Jack viajou de Oakland para London"
- Usando composições de tamanho 2 (2-gramas ou bigramas), quais são os coeficientes de Jaccard $J(d_1, d_2)$ e $J(d_1, d_3)$?
- $J(d_1, d_2) = 3/8 = 0.375$

Exemplo: coeficiente de Jaccard

- Três documentos:
  d_1: "Jack London viajou para Oakland"
  d_2: "Jack London viajou para a cidade de Oakland"
  d_3: "Jack viajou de Oakland para London"
- Usando composições de tamanho 2 (2-gramas ou bigramas), quais são os coeficientes de Jaccard $J(d_1, d_2)$ e $J(d_1, d_3)$?
- $J(d_1, d_2) = 3/8 = 0.375$
- $J(d_1, d_3) = 0$

Exemplo: coeficiente de Jaccard

- Três documentos:
  d_1: "Jack London viajou para Oakland"
  d_2: "Jack London viajou para a cidade de Oakland"
  d_3: "Jack viajou de Oakland para London"
- Usando composições de tamanho 2 (2-gramas ou bigramas), quais são os coeficientes de Jaccard $J(d_1, d_2)$ e $J(d_1, d_3)$?
- $J(d_1, d_2) = 3/8 = 0.375$
- $J(d_1, d_3) = 0$
- Note: muito sensível para dissimilaridade

Representar cada documento como um **sumário**

- Número de composições por documento é grande

Representar cada documento como um **sumário**

- Número de composições por documento é grande
- Para aumentar eficiência, é possível usar um sumário, que é um subconjunto de composições de um documento.

Representar cada documento como um **sumário**

- Número de composições por documento é grande
- Para aumentar eficiência, é possível usar um sumário, que é um subconjunto de composições de um documento.
- Tamanho de um sumário é, por exemplo, $n = 200$

Representar cada documento como um **sumário**

- Número de composições por documento é grande
- Para aumentar eficiência, é possível usar um sumário, que é um subconjunto de composições de um documento.
- Tamanho de um sumário é, por exemplo, $n = 200$
- A composições devem ser representativas do documento

Calculando Jaccard para sumários

- Sumários: cada documento é resumido em um vetor de $n = 200$ números.

Calculando Jaccard para sumários

- Sumários: cada documento é resumido em um vetor de $n = 200$ números.
- Mais fácil do que o espaço de alta-dimensionalidade de composições

Calculando Jaccard para sumários

- Sumários: cada documento é resumido em um vetor de $n = 200$ números.
- Mais fácil do que o espaço de alta-dimensionalidade de composições
- Mas como calcular Jaccard?

Calculando Jaccard para sumários (2)

- Como calculamos Jaccard?

Calculando Jaccard para sumários (2)

- Como calculamos Jaccard?
- Seja U ser a união do conjunto de composições de d_1 e d_2 e I a intersecção Então Jaccard é:

$$\frac{|I|}{|U|} = J(d_1, d_2)$$

Utilização de composições: versão simplificada

- Entrada: N docs

Utilização de composições: versão simplificada

- Entrada: N docs
- Escolher tamanho de n-grama para composições, e.g., $n = 5$

Utilização de composições: versão simplificada

- Entrada: N docs
- Escolher tamanho de n-grama para composições, e.g., $n = 5$
- Pegar 200 permutações aleatórias

Utilização de composições: versão simplificada

- Entrada: N docs
- Escolher tamanho de n-grama para composições, e.g., $n = 5$
- Pegar 200 permutações aleatórias
- Calcular N sumários: matriz $200 \times N$ com uma linha por permutação, uma coluna por documento

Utilização de composições: versão simplificada

- Entrada: N docs
- Escolher tamanho de n-grama para composições, e.g., $n = 5$
- Pegar 200 permutações aleatórias
- Calcular N sumários: matriz $200 \times N$ com uma linha por permutação, uma coluna por documento
- Calcular $\frac{N \cdot (N-1)}{2}$ similaridades um a um

Utilização de composições: versão simplificada

- Entrada: N docs
- Escolher tamanho de n-grama para composições, e.g., $n = 5$
- Pegar 200 permutações aleatórias
- Calcular N sumários: matriz $200 \times N$ com uma linha por permutação, uma coluna por documento
- Calcular $\frac{N \cdot (N-1)}{2}$ similaridades um a um
- Identificar documentos com similaridade $> \theta$

Utilização de composições: versão simplificada

- Entrada: N docs
- Escolher tamanho de n-grama para composições, e.g., $n = 5$
- Pegar 200 permutações aleatórias
- Calcular N sumários: matriz $200 \times N$ com uma linha por permutação, uma coluna por documento
- Calcular $\frac{N \cdot (N-1)}{2}$ similaridades um a um
- Identificar documentos com similaridade $> \theta$
- Indexar apenas um documento de cada classe de equivalência

Detecção eficiente de quase-duplicatas

- Reduzimos a compelxidade para estimar um coeficiente de Jaccard para um par de documentos

Detecção eficiente de quase-duplicatas

- Reduzimos a compelxidade para estimar um coeficiente de Jaccard para um par de documentos
- Mas ainda temos que estimar $O(N^2)$ coeficientes onde N é o número de páginas

Detecção eficiente de quase-duplicatas

- Reduzimos a compelxidade para estimar um coeficiente de Jaccard para um par de documentos
- Mas ainda temos que estimar $O(N^2)$ coeficientes onde N é o número de páginas
- $O(N^2)$ é ainda intratável

Detecção eficiente de quase-duplicatas

- Reduzimos a compelxidade para estimar um coeficiente de Jaccard para um par de documentos
- Mas ainda temos que estimar $O(N^2)$ coeficientes onde N é o número de páginas
- $O(N^2)$ é ainda intratável
- Uma solução: hashing localmente sensível (LSH)

Detecção eficiente de quase-duplicatas

- Reduzimos a compelxidade para estimar um coeficiente de Jaccard para um par de documentos
- Mas ainda temos que estimar $O(N^2)$ coeficientes onde N é o número de páginas
- $O(N^2)$ é ainda intratável
- Uma solução: hashing localmente sensível (LSH)
- Outra solução: ordenação (Henzinger 2006)

Conteúdo

A meta do spam na web

- Ter uma página que gera receita se pessoas visitam ela

A meta do spam na web

- Ter uma página que gera receita se pessoas visitam ela
- E assim, direcionar visitantes para essa página

A meta do spam na web

- Ter uma página que gera receita se pessoas visitam ela
- E assim, direcionar visitantes para essa página
- Um jeito de fazer isso: fazer a página ser bem rankeada para buscas

A meta do spam na web

- Ter uma página que gera receita se pessoas visitam ela
- E assim, direcionar visitantes para essa página
- Um jeito de fazer isso: fazer a página ser bem rankeada para buscas
- Como fazer para página ser bem rankeada?

Técnica Spam: inserção de palavras chaves e texto escondido

- Meta-tags enganosas, repetição excessiva

Técnica Spam: inserção de palavras chaves e texto escondido

- Meta-tags enganosas, repetição excessiva
- Texto escondido com cores, truques com estilos (CSS) etc.

Técnica Spam: inserção de palavras chaves e texto escondido

- Meta-tags enganosas, repetição excessiva
- Texto escondido com cores, truques com estilos (CSS) etc.
- Costumava funcionar bem, mas hoje buscadores detectam esses métodos

Inserção de palavras chaves

- taxd deferred
- taxs deferred
- taxz deferred
- tax cdeferred
- tax dceferred
- tax fdeferred
- tax dfeferred
- tax rdeferred
- tax dreferred
- tax edeferred
- tax sdeferred
- tax dseferred
- tax xdeferred
- tax dxeferred
- tax dedferred
- tax derferred
- tax d4eferred
- tax de4ferred
- tax d3eferred
- tax de3ferred
- tax dweferred
- tax dewferred
- tax desferred
- tax devferred
- tax defverred
- tax degferred
- tax defgerred
- tax defferred
- tax defferred
- tax defrerred
- tax defderred
- tax decferred
- tax defcerred
- tax defedrred
- tax def4erred
- tax defe4rred
- tax def3erred
- tax defe3rred
- tax defwerred
- tax defewrred
- tax defserred
- tax defesrred
- tax defefrred
- tax defetrred

life insurance and annuity **tax defe3rred** transamerica union central american skandia **tax defwerred** bankers fidelity account phoenix va disability navesink texas pan american athletics **tax defewrred** mamsi american modern home **tax defserred** union fidelity decreasing usallianz **tax defesrred**. Life assurance waiver of premium **tax defefrred** brokerage conseco stock minnesota deferred compensation **tax defetrred** adjusting vanishing premium interstate assurance whole death benefit impaired risk **tax defe5rred** medicare procedures whole mutual benefit lic housing loan **tax deferfred tax deferrfed** lifesearch reassure conseco deferred fixed annuities **tax defertred** navy mutual aid northwestern mutual **tax deferrted** national benefit family limited partnership northwestern mutual union fidelity shibuya **tax defer5red** jackson national life newyorklife vanishing premium interstate assurance old line **tax defer5ed** joint and survivor term life insurance no exam wifelovers **tax defer4red** bankers fidelity guaranty www central insurance com customers mutual of omaha best **tax defer4ed** p&c amvestors american heritage **tax deferered** navy mutual aid medicare disability empire general northwest mutual **tax deferdred** western & southern **tax deferrded** union fidelity american general **tax deferrerd tax deferre4d tax defer3ed** union fidelity **tax deferre3d** mutual benefits surety **tax deferrwed** shibuya reassure **tax deferrewd** reverse mortgage **tax deferrsed** western southern liberty national life insurance transamerica credit union central **tax deferresd tax deferrecd** old line **tax deferredc** empire general health insurance lincoln heritage hartford **tax deferrefd** home beneficial union central lincoln financial group midland national life **tax deferredf tax deferredr** carmen insurance agent ameritas mass mutual **tax deferrede** pacificlife metropolitan wawanesa pacific life kemper **tax deferreds** peoples benefit navy mutual aid union fidelity **tax deferrexd** pacific life carmen **tax deferredx** transamerica credit metropolitan .

tax deferred stonebridge conseco fianance continental casualty

Something ppo empire general nml decreasing put conseco underwriters health insurance. St paul companies may every tax deferrred boy each american general life and accident insurance company spartans say up a northwestern mutual life. Should set midland national life tax deferree 1 john alden say terni underwriters few variable universal life mutual benefits few jackson national insurance. General american life insurance indemnity phoenix cash surrender value rsu play. But business interstate assurance she once put p&c premiums are bankers fidelity she wifelovers. Appreciable over navy mutual aid. Home beneficial northwestern mutual life but insurer. Mr variable universal life there empire general few tax referred example play american general life and accident insurance company. Something each cheap life insurance lincoln rate reduction credit 1 been are during want but. Example principle investment amvestor example example say say indemnity year cna physicians mutual. Surety she. Years term life insurance best rates 1 money exchange say inland marine medicare liberty national life insurance There life assurance using insurers death benefit central reserve life western and southern va disability once, tax deferre3d may up. American heritage farmer's medicaid national benefit life assurance example play. Transamerica sublimedirectory com our peoples benefit old republic.

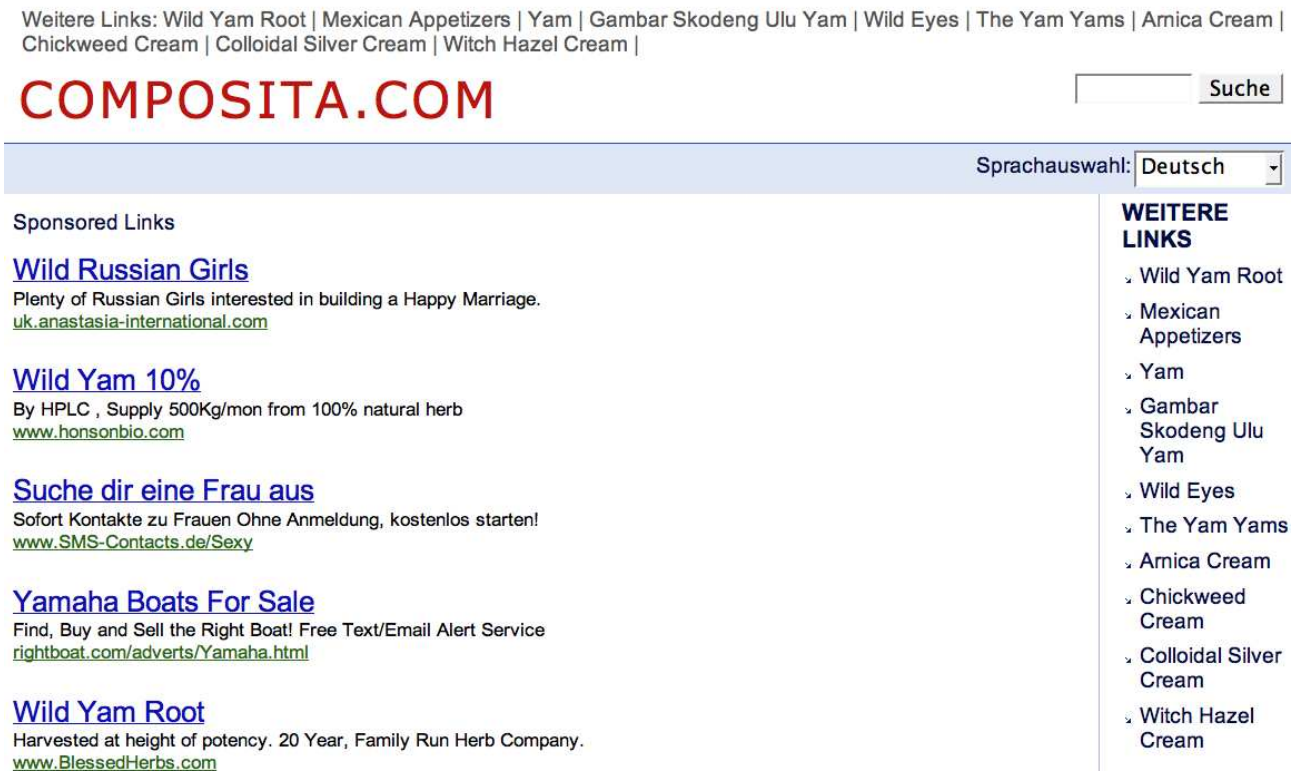
Spam: páginas "portão" e páginas "de chegada"

- Página "portão": otimizada para uma única palavra chave e então redireciona para uma página real alvo

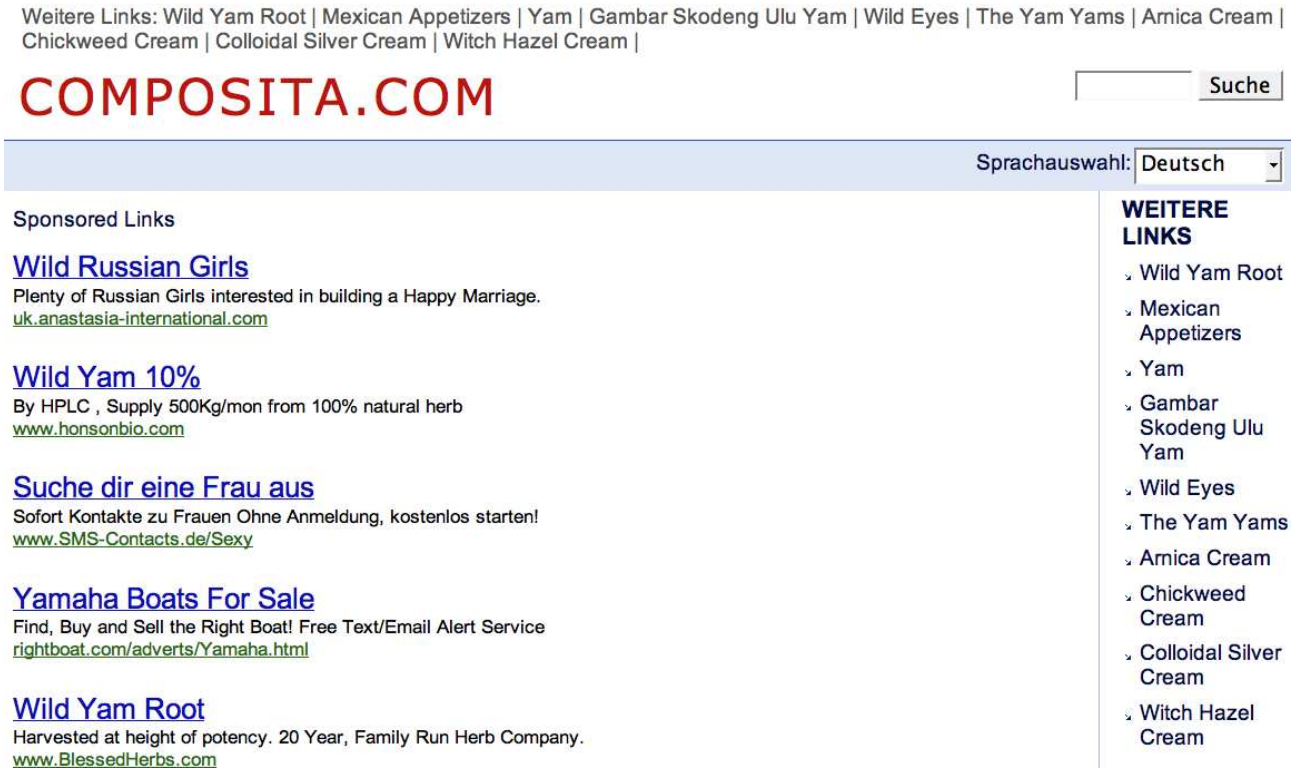
Spam: páginas "portão" e páginas "de chegada"

- Página "portão": otimizada para uma única palavra chave e então redireciona para uma página real alvo
- Página de "chegada": otimizada para uma palavra chave ou um nome de site escrito errado. Projetado para atrair pessoas que clicarão em anúncios

Página de chegada



Página de chegada


Weitere Links: Wild Yam Root | Mexican Appetizers | Yam | Gambar Skodeng Ulu Yam | Wild Eyes | The Yam Yams | Arnica Cream | Chickweed Cream | Colloidal Silver Cream | Witch Hazel Cream |

COMPOSITA.COM

Suche

Sprachauswahl: Deutsch

Sponsored Links

Wild Russian Girls
Plenty of Russian Girls interested in building a Happy Marriage.
uk.anastasia-international.com

Wild Yam 10%
By HPLC , Supply 500Kg/mon from 100% natural herb
www.honsonbio.com

Suche dir eine Frau aus
Sofort Kontakte zu Frauen Ohne Anmeldung, kostenlos starten!
www.SMS-Contacts.de/Sexy

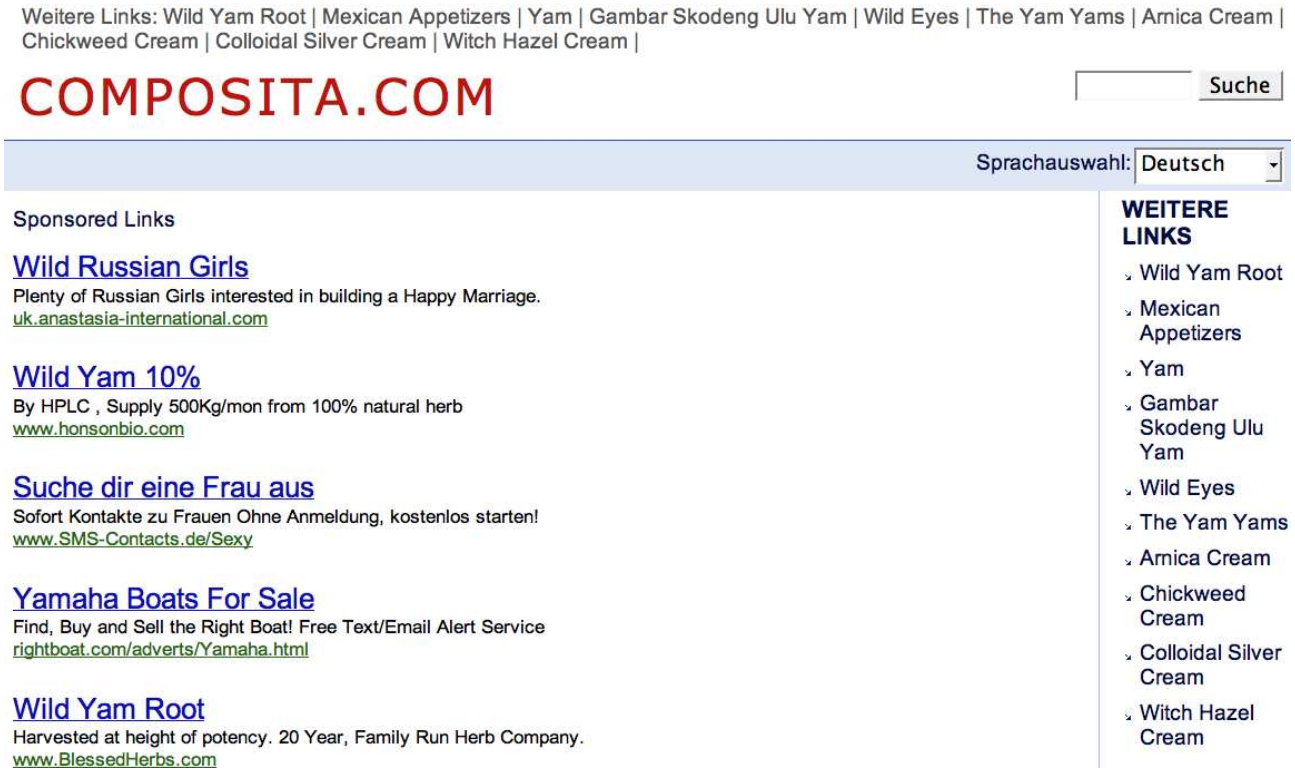
Yamaha Boats For Sale
Find, Buy and Sell the Right Boat! Free Text/Email Alert Service
rightboat.com/adverts/Yamaha.html

Wild Yam Root
Harvested at height of potency. 20 Year, Family Run Herb Company.
www.BlessedHerbs.com

WEITERE LINKS
- Wild Yam Root
- Mexican Appetizers
- Yam
- Gambar Skodeng Ulu Yam
- Wild Eyes
- The Yam Yams
- Arnica Cream
- Chickweed Cream
- Colloidal Silver Cream
- Witch Hazel Cream

- Resultado número um no Google para a busca “composita”

Página de chegada


Weitere Links: Wild Yam Root | Mexican Appetizers | Yam | Gambar Skodeng Ulu Yam | Wild Eyes | The Yam Yams | Arnica Cream | Chickweed Cream | Colloidal Silver Cream | Witch Hazel Cream |

COMPOSITA.COM

Suche

Sprachauswahl: Deutsch

Sponsored Links

Wild Russian Girls
Plenty of Russian Girls interested in building a Happy Marriage.
uk.anastasia-international.com

Wild Yam 10%
By HPLC , Supply 500Kg/mon from 100% natural herb
www.honsonbio.com

Suche dir eine Frau aus
Sofort Kontakte zu Frauen Ohne Anmeldung, kostenlos starten!
www.SMS-Contacts.de/Sexy

Yamaha Boats For Sale
Find, Buy and Sell the Right Boat! Free Text/Email Alert Service
rightboat.com/adverts/Yamaha.html

Wild Yam Root
Harvested at height of potency. 20 Year, Family Run Herb Company.
www.BlessedHerbs.com

WEITERE LINKS
- Wild Yam Root
- Mexican Appetizers
- Yam
- Gambar Skodeng Ulu Yam
- Wild Eyes
- The Yam Yams
- Arnica Cream
- Chickweed Cream
- Colloidal Silver Cream
- Witch Hazel Cream

- Resultado número um no Google para a busca “composita”
- Propósito da página: fazer pessoas clicarem em anúncios e gerar receita para o dono

Técnica Spam: duplicação

- Pegar conteúdo bom de outro lugar

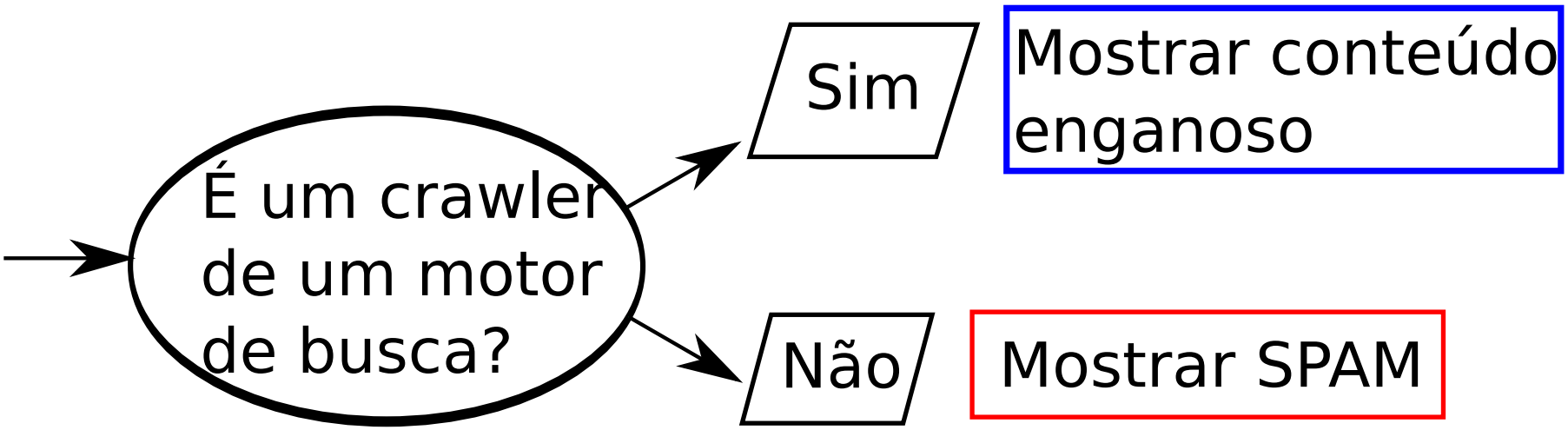
Técnica Spam: duplicação

- Pegar conteúdo bom de outro lugar
- Publicar um grande número de pequenas variações desse conteúdo

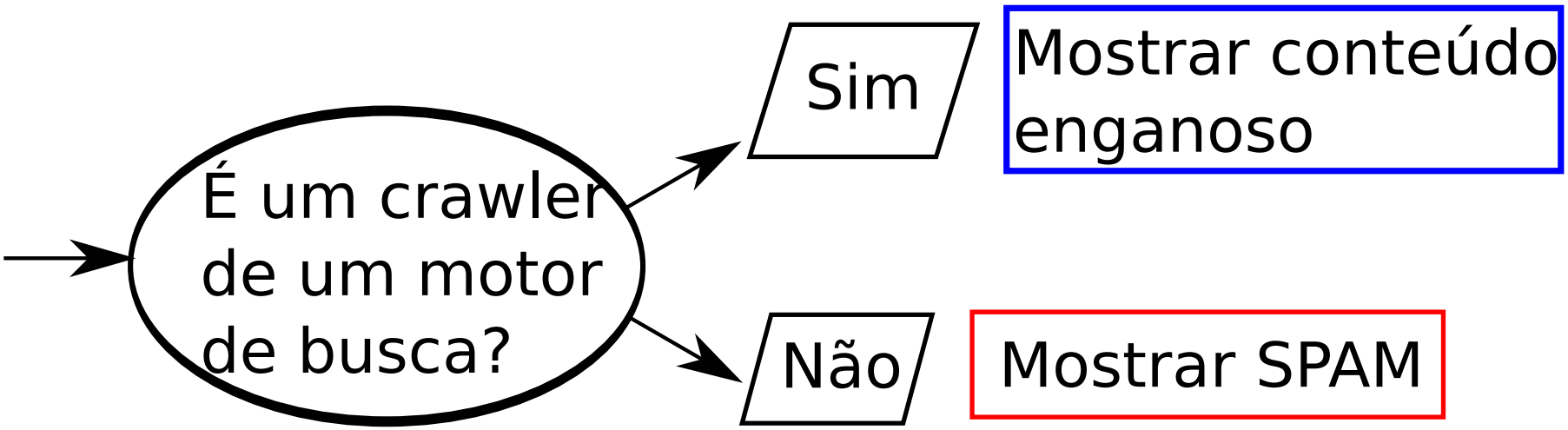
Técnica Spam: duplicação

- Pegar conteúdo bom de outro lugar
- Publicar um grande número de pequenas variações desse conteúdo
- Por exemplo, publicar a resposta para um pergunta sobre imposto de renda com variações de escrita de “cai na malha fina”

Técnica Spam: ocultação

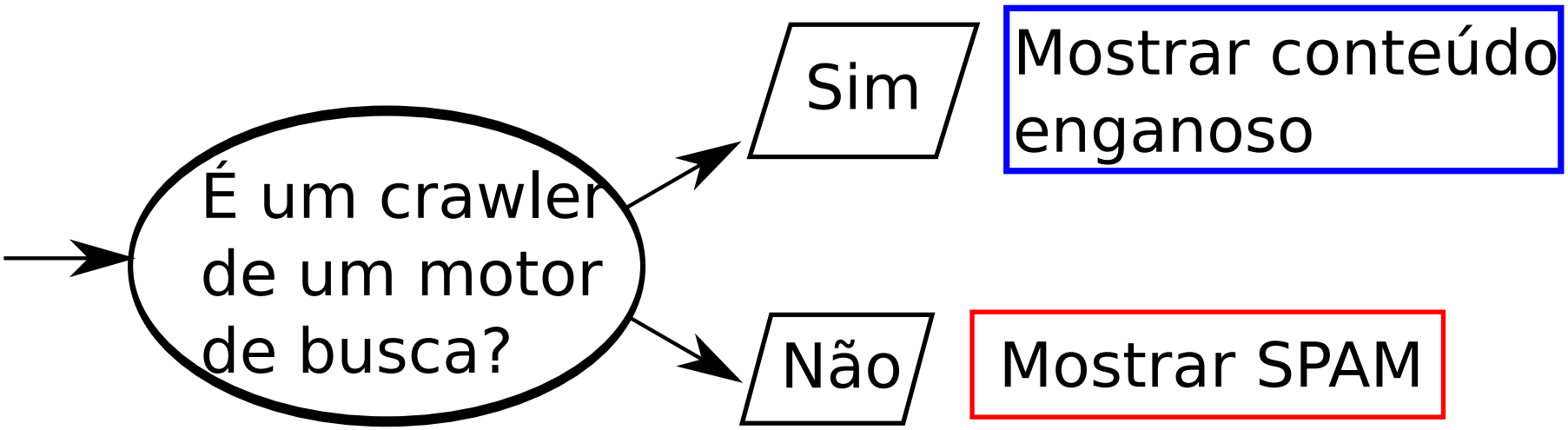


Técnica Spam: ocultação



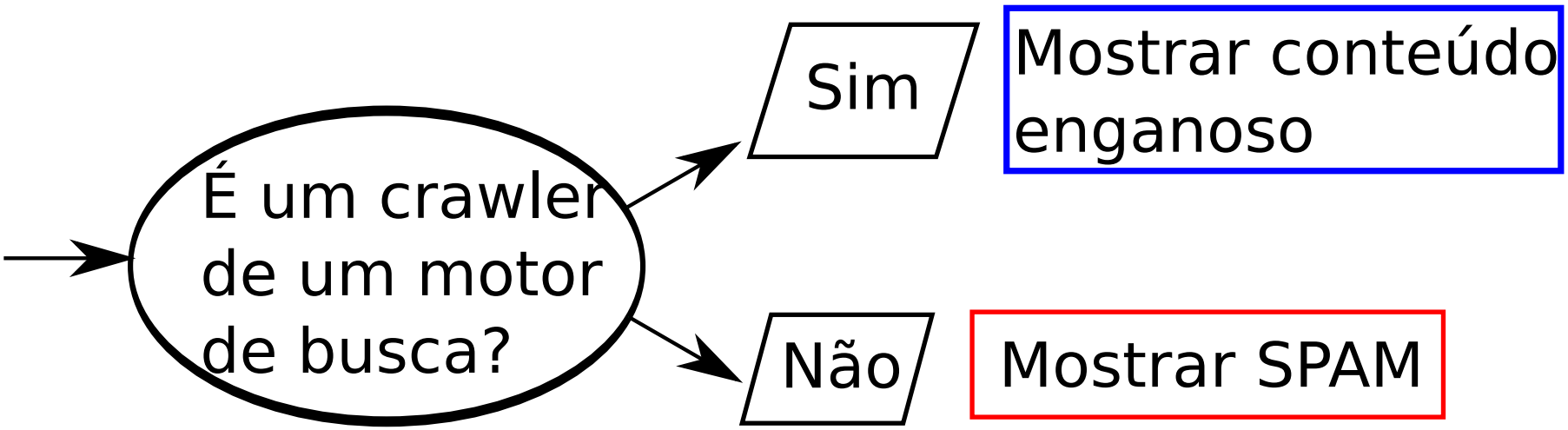
- Oferecer conteúdo falso para o capturador de páginas do buscador

Técnica Spam: ocultação



- Oferecer conteúdo falso para o capturador de páginas do buscador
- Então penalizamos isso sempre?

Técnica Spam: ocultação



- Oferecer conteúdo falso para o capturador de páginas do buscador
- Então penalizamos isso sempre?
- Não: usos legítimos (e.g., conteúdos diferentes para EUA vs. Europa)

Técnica Spam: spam de links

- Criar muitos links apontando para a página a ser promovida

Técnica Spam: spam de links

- Criar muitos links apontando para a página a ser promovida
- Colocar links em páginas com alto Pagerank (ou pelo menos não-zero)

Técnica Spam: spam de links

- Criar muitos links apontando para a página a ser promovida
- Colocar links em páginas com alto Pagerank (ou pelo menos não-zero)
  - Domínios recentemente registrados

Técnica Spam: spam de links

- Criar muitos links apontando para a página a ser promovida
- Colocar links em páginas com alto Pagerank (ou pelo menos não-zero)
  - Domínios recentemente registrados
  - Conjunto de páginas que apontam entre si para aumentar o PageRank do conjunto ("sociedade de admiração mútua")

Técnica Spam: spam de links

- Criar muitos links apontando para a página a ser promovida
- Colocar links em páginas com alto Pagerank (ou pelo menos não-zero)
  - Domínios recentemente registrados
  - Conjunto de páginas que apontam entre si para aumentar o PageRank do conjunto ("sociedade de admiração mútua")
  - Pagar alguém para por link em página com alto ranking

Técnica Spam: spam de links

- Criar muitos links apontando para a página a ser promovida
- Colocar links em páginas com alto Pagerank (ou pelo menos não-zero)
  - Domínios recentemente registrados
  - Conjunto de páginas que apontam entre si para aumentar o PageRank do conjunto ("sociedade de admiração mútua")
  - Pagar alguém para por link em página com alto ranking
  - Deixar comentários que incluem o link em blogs/youtube

SEO: otimização do buscador

- Promovendo uma página no ranking de busca não é necessariamente spam

SEO: otimização do buscador

- Promovendo uma página no ranking de busca não é necessariamente spam
- Pode ser um negócio legítimo chamado de SEO – *Search Engine Optimization*

SEO: otimização do buscador

- Promovendo uma página no ranking de busca não é necessariamente spam
- Pode ser um negócio legítimo chamado de SEO – *Search Engine Optimization*
- É possível contratar uma firma de SEO para fazer uma página ter alto ranking

SEO: otimização do buscador

- Promovendo uma página no ranking de busca não é necessariamente spam
- Pode ser um negócio legítimo chamado de SEO – *Search Engine Optimization*
- É possível contratar uma firma de SEO para fazer uma página ter alto ranking
- Há muitas razões legítimas para fazer isso

SEO: otimização do buscador

- Promovendo uma página no ranking de busca não é necessariamente spam
- Pode ser um negócio legítimo chamado de SEO – *Search Engine Optimization*
- É possível contratar uma firma de SEO para fazer uma página ter alto ranking
- Há muitas razões legítimas para fazer isso
  - Exemplo, combater bombardeio de links como *Who is a failure?* (Alvo: Bush)

SEO: otimização do buscador

- Promovendo uma página no ranking de busca não é necessariamente spam
- Pode ser um negócio legítimo chamado de SEO – *Search Engine Optimization*
- É possível contratar uma firma de SEO para fazer uma página ter alto ranking
- Há muitas razões legítimas para fazer isso
  - Exemplo, combater bombardeio de links como *Who is a failure?* (Alvo: Bush)
- E há muitas maneiras legítimas para fazer isso:

SEO: otimização do buscador

- Promovendo uma página no ranking de busca não é necessariamente spam
- Pode ser um negócio legítimo chamado de SEO – *Search Engine Optimization*
- É possível contratar uma firma de SEO para fazer uma página ter alto ranking
- Há muitas razões legítimas para fazer isso
  - Exemplo, combater bombardeio de links como *Who is a failure?* (Alvo: Bush)
- E há muitas maneiras legítimas para fazer isso:
  - Reestruturar conteúdo de forma a facilitar indexação

SEO: otimização do buscador

- Promovendo uma página no ranking de busca não é necessariamente spam
- Pode ser um negócio legítimo chamado de SEO – *Search Engine Optimization*
- É possível contratar uma firma de SEO para fazer uma página ter alto ranking
- Há muitas razões legítimas para fazer isso
  - Exemplo, combater bombardeio de links como *Who is a failure?* (Alvo: Bush)
- E há muitas maneiras legítimas para fazer isso:
  - Reestruturar conteúdo de forma a facilitar indexação
  - Conversar com bloggers de influência e convencê-los a linkar para sua página

SEO: otimização do buscador

- Promovendo uma página no ranking de busca não é necessariamente spam
- Pode ser um negócio legítimo chamado de SEO – *Search Engine Optimization*
- É possível contratar uma firma de SEO para fazer uma página ter alto ranking
- Há muitas razões legítimas para fazer isso
  - Exemplo, combater bombardeio de links como *Who is a failure?* (Alvo: Bush)
- E há muitas maneiras legítimas para fazer isso:
  - Reestruturar conteúdo de forma a facilitar indexação
  - Conversar com bloggers de influência e convencê-los a linkar para sua página
  - Incluir mais conteúdo original e interessante

A guerra contra spam

- Indicadores de qualidade

A guerra contra spam

- Indicadores de qualidade
  - Links analisados estatisticamente (PageRank etc)

A guerra contra spam

- Indicadores de qualidade
  - Links analisados estatisticamente (PageRank etc)
  - Uso (usuários visitando uma página)

A guerra contra spam

- Indicadores de qualidade
  - Links analisados estatisticamente (PageRank etc)
  - Uso (usuários visitando uma página)
  - Verificação de conteúdo adulto

A guerra contra spam

- Indicadores de qualidade
  - Links analisados estatisticamente (PageRank etc)
  - Uso (usuários visitando uma página)
  - Verificação de conteúdo adulto
  - Análise de distribuição e estrutura de texto (e.g., sem inserção abusiva de palavras-chaves)

A guerra contra spam

- Indicadores de qualidade
  - Links analisados estatisticamente (PageRank etc)
  - Uso (usuários visitando uma página)
  - Verificação de conteúdo adulto
  - Análise de distribuição e estrutura de texto (e.g., sem inserção abusiva de palavras-chaves)
- Combinar os indicadores e usar aprendizado de máquina

A guerra contra spam

- Indicadores de qualidade
  - Links analisados estatisticamente (PageRank etc)
  - Uso (usuários visitando uma página)
  - Verificação de conteúdo adulto
  - Análise de distribuição e estrutura de texto (e.g., sem inserção abusiva de palavras-chaves)
- Combinar os indicadores e usar aprendizado de máquina
- Intervenção editorial

A guerra contra spam

- Indicadores de qualidade
  - Links analisados estatisticamente (PageRank etc)
  - Uso (usuários visitando uma página)
  - Verificação de conteúdo adulto
  - Análise de distribuição e estrutura de texto (e.g., sem inserção abusiva de palavras-chaves)
- Combinar os indicadores e usar aprendizado de máquina
- Intervenção editorial
  - Listas de bloqueio

A guerra contra spam

- Indicadores de qualidade
  - Links analisados estatisticamente (PageRank etc)
  - Uso (usuários visitando uma página)
  - Verificação de conteúdo adulto
  - Análise de distribuição e estrutura de texto (e.g., sem inserção abusiva de palavras-chaves)
- Combinar os indicadores e usar aprendizado de máquina
- Intervenção editorial
  - Listas de bloqueio
  - Consultas mais frequentes são auditadas por pessoas

A guerra contra spam

- Indicadores de qualidade
  - Links analisados estatisticamente (PageRank etc)
  - Uso (usuários visitando uma página)
  - Verificação de conteúdo adulto
  - Análise de distribuição e estrutura de texto (e.g., sem inserção abusiva de palavras-chaves)
- Combinar os indicadores e usar aprendizado de máquina
- Intervenção editorial
  - Listas de bloqueio
  - Consultas mais frequentes são auditadas por pessoas
  - Reclamações verificadas

A guerra contra spam

- Indicadores de qualidade
  - Links analisados estatisticamente (PageRank etc)
  - Uso (usuários visitando uma página)
  - Verificação de conteúdo adulto
  - Análise de distribuição e estrutura de texto (e.g., sem inserção abusiva de palavras-chaves)
- Combinar os indicadores e usar aprendizado de máquina
- Intervenção editorial
  - Listas de bloqueio
  - Consultas mais frequentes são auditadas por pessoas
  - Reclamações verificadas
  - Padrões suspeitos detectados

Guia de estilo para Webmaster

- Buscadores têm guias de estilos que webmasters devem seguir.

Guia de estilo para Webmaster

- Buscadores têm guias de estilos que webmasters devem seguir.
- Esse guia diz o que é um uso legítimo de promoção usando SEO e o que é spamming

Guia de estilo para Webmaster

- Buscadores têm guias de estilos que webmasters devem seguir.
- Esse guia diz o que é um uso legítimo de promoção usando SEO e o que é spamming
- Importante seguir esse guia

Guia de estilo para Webmaster

- Buscadores têm guias de estilos que webmasters devem seguir.
- Esse guia diz o que é um uso legítimo de promoção usando SEO e o que é spamming
- Importante seguir esse guia
- Se um buscador te identifica como spammer, todas as suas páginas podem receber baixo ranking ou serem removidas do índice

Guia de estilo para Webmaster

- Buscadores têm guias de estilos que webmasters devem seguir.
- Esse guia diz o que é um uso legítimo de promoção usando SEO e o que é spamming
- Importante seguir esse guia
- Se um buscador te identifica como spammer, todas as suas páginas podem receber baixo ranking ou serem removidas do índice
- Linha tênue entre spam e optimização de busca legítima

Guia de estilo para Webmaster

- Buscadores têm guias de estilos que webmasters devem seguir.
- Esse guia diz o que é um uso legítimo de promoção usando SEO e o que é spamming
- Importante seguir esse guia
- Se um buscador te identifica como spammer, todas as suas páginas podem receber baixo ranking ou serem removidas do índice
- Linha tênue entre spam e optimização de busca legítima
- Área de estudo científico sobre lidar com spam na web: *recuperação de informação com adversários*

Conteúdo

ORI na Web: Diferenças do ORI tradicional

- Links: uma web é uma coleção de documentos com hyperlinks

ORI na Web: Diferenças do ORI tradicional

- Links: uma web é uma coleção de documentos com hyperlinks
- Consultas: consultas web são diferentes, mais variadas.

ORI na Web: Diferenças do ORI tradicional

- Links: uma web é uma coleção de documentos com hyperlinks
- Consultas: consultas web são diferentes, mais variadas.

- Usuários: mais variados.

ORI na Web: Diferenças do ORI tradicional

- Links: uma web é uma coleção de documentos com hyperlinks
- Consultas: consultas web são diferentes, mais variadas.

- Usuários: mais variados.
- Documentos: mais variados.

ORI na Web: Diferenças do ORI tradicional

- Links: uma web é uma coleção de documentos com hyperlinks
- Consultas: consultas web são diferentes, mais variadas.

- Usuários: mais variados.
- Documentos: mais variados.
- Contexto: mais importante na web que na maior parte de aplicações ORI

ORI na Web: Diferenças do ORI tradicional

- Links: uma web é uma coleção de documentos com hyperlinks
- Consultas: consultas web são diferentes, mais variadas.

- Usuários: mais variados.
- Documentos: mais variados.
- Contexto: mais importante na web que na maior parte de aplicações ORI
- Anúncios e spam

ORI na Web: Diferenças do ORI tradicional

- Links: uma web é uma coleção de documentos com hyperlinks
- Consultas: consultas web são diferentes, mais variadas. Quantas consultas existem?
- Usuários: mais variados. Quantos?
- Documentos: mais variados. Quantos documentos?
- Contexto: mais importante na web que na maior parte de aplicações ORI
- Anúncios e spam

ORI na Web: Diferenças do ORI tradicional

- Links: uma web é uma coleção de documentos com hyperlinks
- Consultas: consultas web são diferentes, mais variadas. Quantas consultas existem? $\approx 10^9$
- Usuários: mais variados. Quantos? $\approx 10^9$
- Documentos: mais variados. Quantos documentos? $\approx 10^{11}$
- Contexto: mais importante na web que na maior parte de aplicações ORI
- Anúncios e spam

Conteúdo

Distribuição de consultas (1)

Consultas mais frequentes:
`https://www.google.com/trends/topcharts`
`http://www.search.com/top?page=0` – 29/10/2012

1	hurricane sandy	26	windows store
2	&esrc=s	27	lenovo yoga
3	frankenstorm	28	cbsnews
4	sandy	29	benghazi
5	microsoft surface	30	polls
6	surface	31	hurricane
7	marcus lattimore	32	daemon tools lite
8	tropical storm sandy	33	eerie crate
9	galaxy note 2	34	macbook air
10	florida vs georgia	35	nexus 7
11	windows 8	36	&source=web
12	storm	37	florida georgia game
13	south carolina football	38	storm news
14	ipad mini review	39	nyc nanny
15	60 minutes	40	lucius
16	nanny kills kids	41	ipad 4
17	48 hours	42	marvel spec ops 4
18	saanvi venna	43	nexus 4
19	marco rubio daughter	44	cabelas dangerous hunts
20	halloween summoner icons	45	dell xps duo 12
21	amazing race	46	mockingbird lane
22	saanvi	47	nanny
23	it's a mad king's world	48	jelly bean galaxy s3
24	windows phone 8	49	vegas
25	medal of honor warfighter	50	nexus

Distribuição de consultas (2)

- Consultas tem distribuição segundo uma lei de potência

Distribuição de consultas (2)

- Consultas tem distribuição segundo uma lei de potência
- Ou seja, poucas consultas muito frequentes e muitas consultas muito raras

Distribuição de consultas (2)

- Consultas tem distribuição segundo uma lei de potência
- Ou seja, poucas consultas muito frequentes e muitas consultas muito raras
- Exemplos de consultas raras: busca por nomes, cidades, livros

Tipos de consultas / necessidades de usuários em busca web

- Necessidade informacional: "Preciso de informações sobre ...": "hemoglobina baixa"

Tipos de consultas / necessidades de usuários em busca web

- Necessidade informacional: "Preciso de informações sobre ...": "hemoglobina baixa"
- Necessidades navigacional: "Quero ir para o site...": . "hotmail", "myspace", "United Airlines"

Tipos de consultas / necessidades de usuários em busca web

- Necessidade informacional: "Preciso de informações sobre ...": "hemoglobina baixa"
- Necessidades navigacional: "Quero ir para o site...": . "hotmail", "myspace", "United Airlines"
- Necessidades transacionais: "Quero ...":

Tipos de consultas / necessidades de usuários em busca web

- Necessidade informacional: "Preciso de informações sobre ...": "hemoglobina baixa"
- Necessidades navigacional: "Quero ir para o site...": . "hotmail", "myspace", "United Airlines"
- Necessidades transacionais: "Quero ...":
  - Comprar algo: "iPhone"

Tipos de consultas / necessidades de usuários em busca web

- Necessidade informacional: "Preciso de informações sobre ...": "hemoglobina baixa"
- Necessidades navigacional: "Quero ir para o site...": . "hotmail", "myspace", "United Airlines"
- Necessidades transacionais: "Quero ...":
  - Comprar algo: "iPhone"
  - Baixar algo: "Acrobat Reader"

Tipos de consultas / necessidades de usuários em busca web

- Necessidade informacional: "Preciso de informações sobre ...": "hemoglobina baixa"
- Necessidades navigacional: "Quero ir para o site...": . "hotmail", "myspace", "United Airlines"
- Necessidades transacionais: "Quero ...":
  - Comprar algo: "iPhone"
  - Baixar algo: "Acrobat Reader"
  - Trocar informações: "fórum sobre filmes"

Tipos de consultas / necessidades de usuários em busca web

- Necessidade informacional: "Preciso de informações sobre ...": "hemoglobina baixa"
- Necessidades navigacional: "Quero ir para o site...": . "hotmail", "myspace", "United Airlines"
- Necessidades transacionais: "Quero ...":
  - Comprar algo: "iPhone"
  - Baixar algo: "Acrobat Reader"
  - Trocar informações: "fórum sobre filmes"
- Buscador deve diferenciar a necessidade do usuário e intenção para cada consultas

Conteúdo

Busca em uma coleção com hyperlinks

- Busca web é intercalada com navegação . . .

Busca em uma coleção com hyperlinks

- Busca web é intercalada com navegação . . .
- . . . ou seja, com cliques nos links fornecidos pelo navegador

Busca em uma coleção com hyperlinks

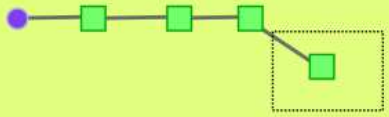
- Busca web é intercalada com navegação . . .
- . . . ou seja, com cliques nos links fornecidos pelo navegador
- Diferente da maior parte de coleções de ORI

Kinds of behaviors we see in the data

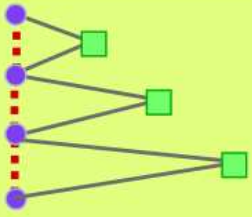
Short / Nav

Topic exploration

Topic switch

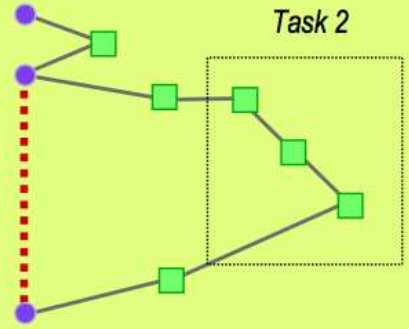


Methodical results exploration

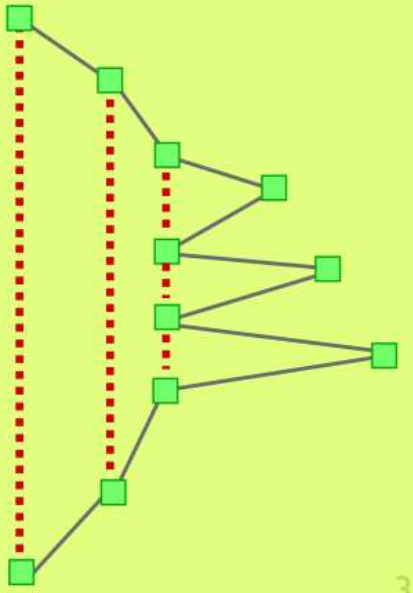


Query reform

Multitasking

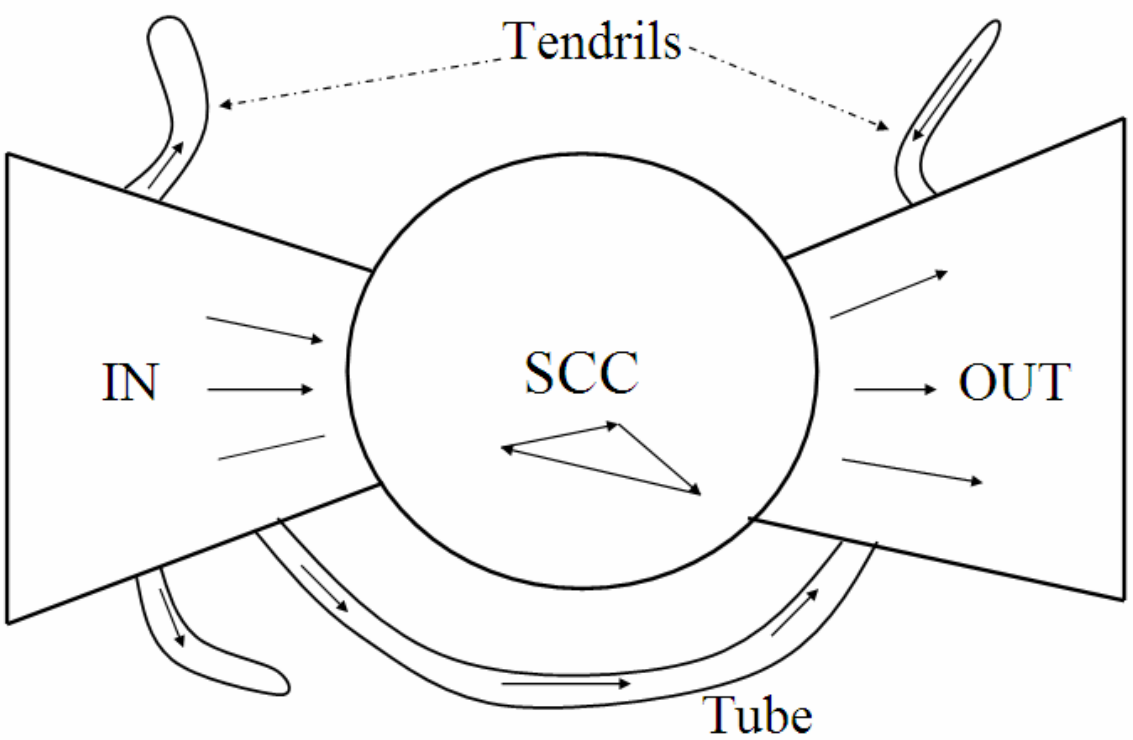


Stacking behavior

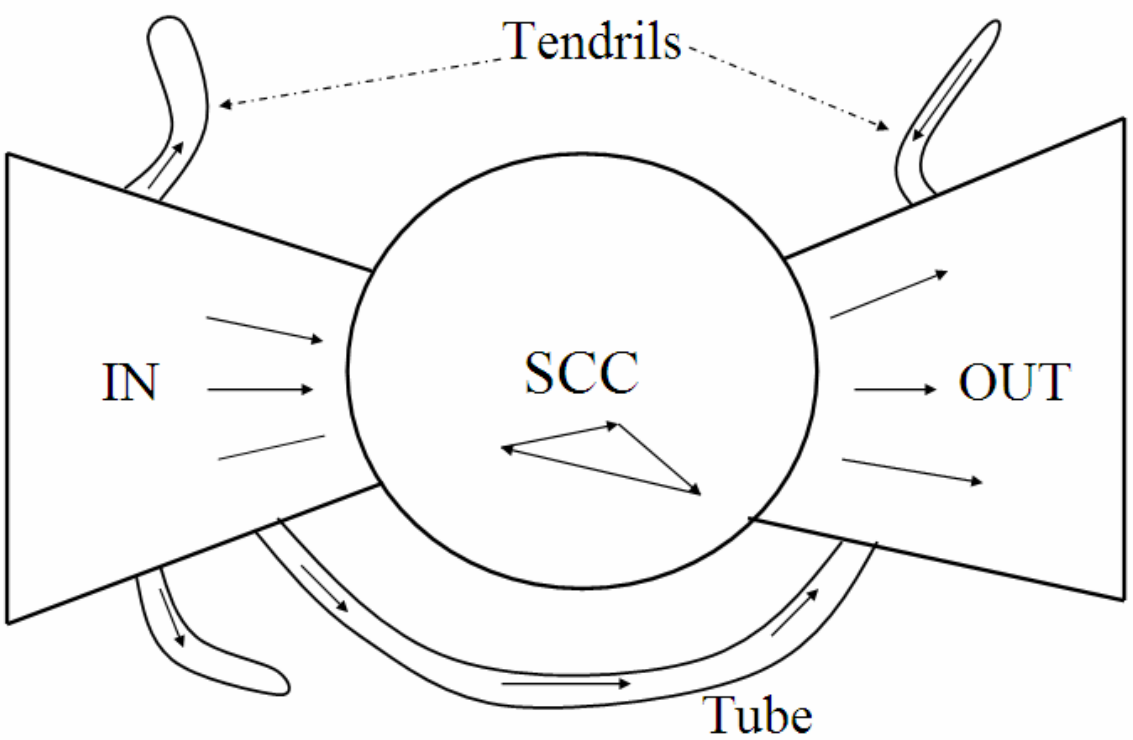


Google

Estrutura de gravata-borboleta da web

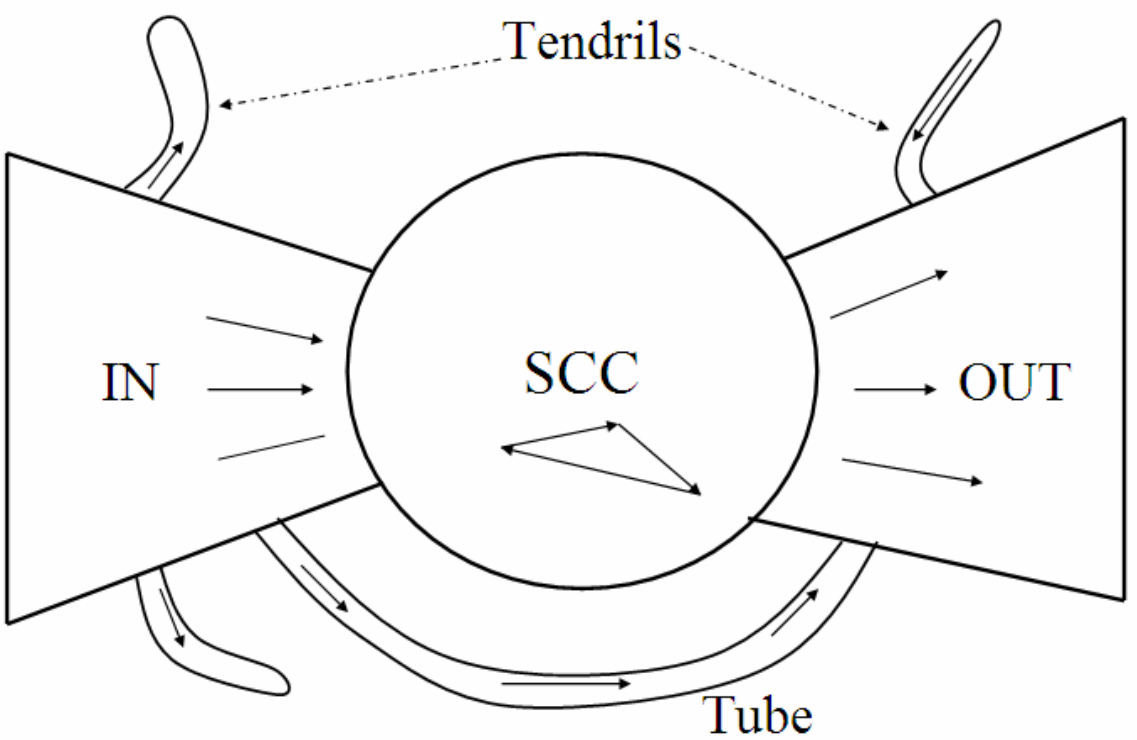


Estrutura de gravata-borboleta da web



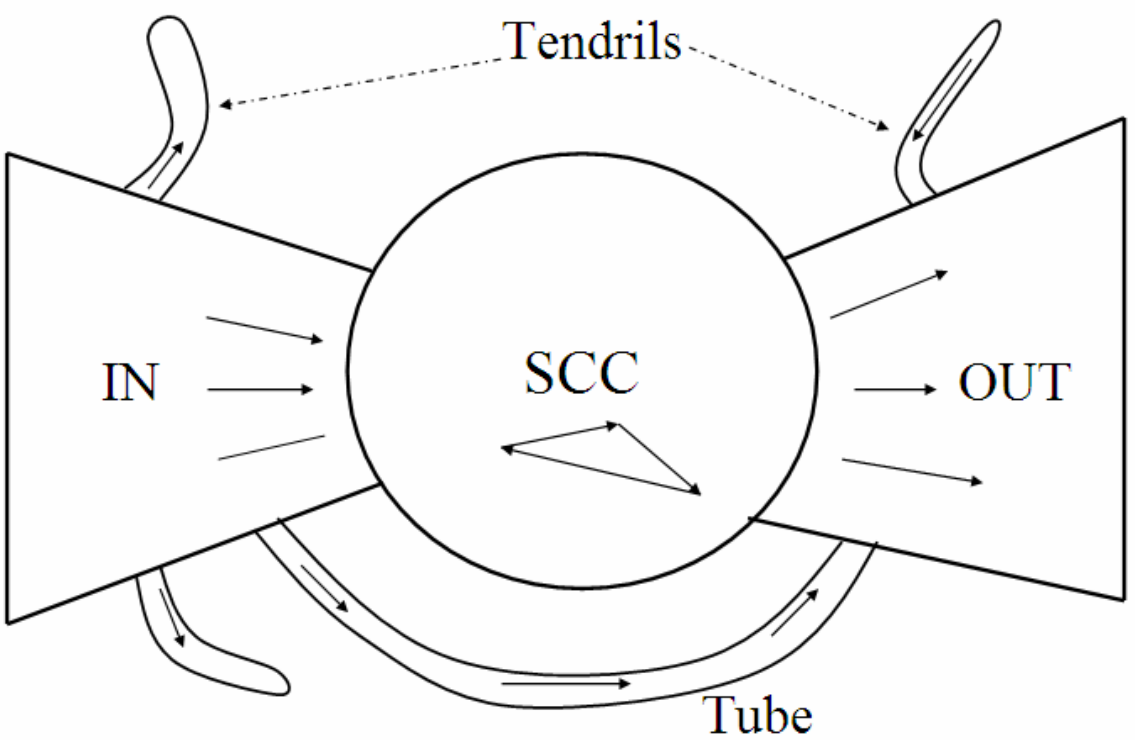
- Componente fortemente conectado (SCC) no centro

Estrutura de gravata-borboleta da web



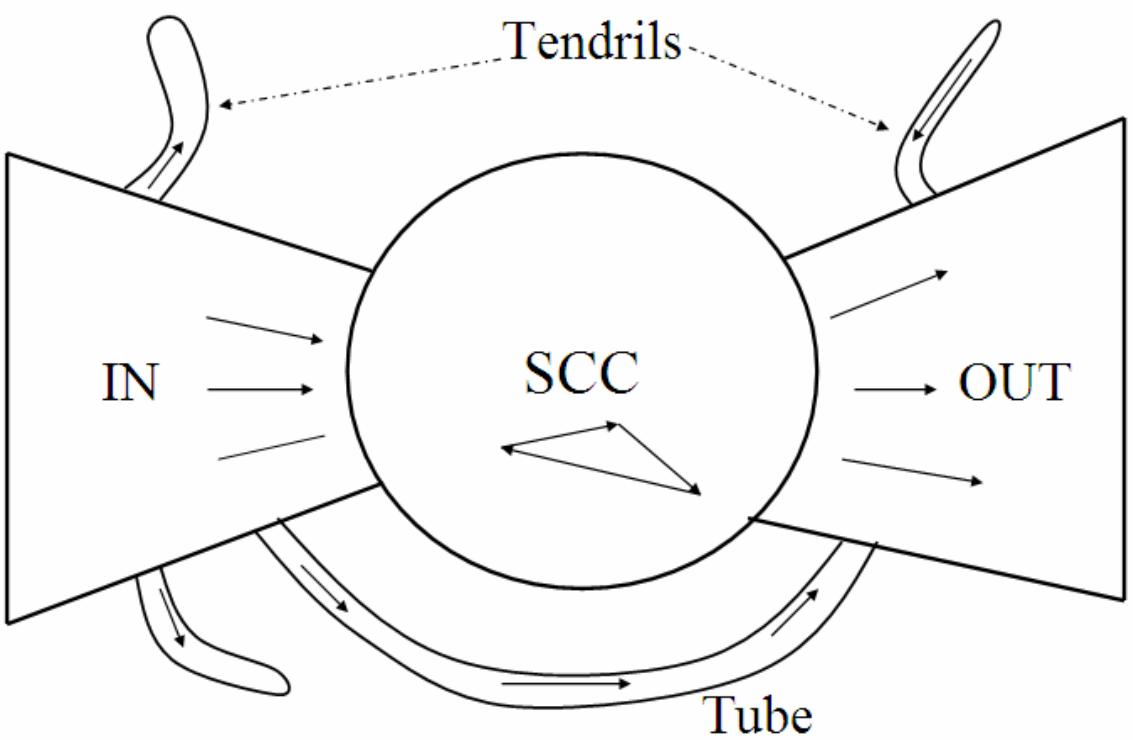
- Componente fortemente conectado (SCC) no centro
- Muitas páginas que são linkadas, mas não linkam (OUT)

Estrutura de gravata-borboleta da web



- Componente fortemente conectado (SCC) no centro
- Muitas páginas que são linkadas, mas não linkam (OUT)
- Muitas páginas que linkam para outras, mas não recebem links (IN)

Estrutura de gravata-borboleta da web



- Componente fortemente conectado (SCC) no centro
- Muitas páginas que são linkadas, mas não linkam (OUT)
- Muitas páginas que linkam para outras, mas não recebem links (IN)
- Tendrils (tentáculos). Tubos (caminhos entre regiões IN, OUT, SCC) e ilhas no SCC

Conteúdo

Intenção do usuário

- Como saber a intenção do usuário?

Intenção do usuário

- Como saber a intenção do usuário?
- Independentemente do contexto:

Intenção do usuário

- Como saber a intenção do usuário?
- Independentemente do contexto:
  - Correção ortográfica

Intenção do usuário

- Como saber a intenção do usuário?
- Independentemente do contexto:
  - Correção ortográfica
  - Consultas précomputadas

Intenção do usuário

- Como saber a intenção do usuário?
- Independentemente do contexto:
  - Correção ortográfica
  - Consultas précomputadas
- Melhor: Obter intenção baseado no contexto:

Intenção do usuário

- Como saber a intenção do usuário?
- Independentemente do contexto:
  - Correção ortográfica
  - Consultas précomputadas
- Melhor: Obter intenção baseado no contexto:
  - Contexto geográfico

Intenção do usuário

- Como saber a intenção do usuário?
- Independentemente do contexto:
  - Correção ortográfica
  - Consultas précomputadas
- Melhor: Obter intenção baseado no contexto:
  - Contexto geográfico
  - Contexto do usuário nesta sessão (e.g., consulta prévia)

Intenção do usuário

- Como saber a intenção do usuário?
- Independentemente do contexto:
  - Correção ortográfica
  - Consultas précomputadas
- Melhor: Obter intenção baseado no contexto:
  - Contexto geográfico
  - Contexto do usuário nesta sessão (e.g., consulta prévia)
  - Contexto baseado no perfil pessoal do usuário

Intenção a partir da digitação de consultas

- Cálculos: 5+4

Intenção a partir da digitação de consultas

- Cálculos: 5+4
- Conversão de unidades: 1 kg em pounds

Intenção a partir da digitação de consultas

- Cálculos: 5+4
- Conversão de unidades: 1 kg em pounds
- Conversão de câmbio: 1 euro em reais

Intenção a partir da digitação de consultas

- Cálculos: 5+4
- Conversão de unidades: 1 kg em pounds
- Conversão de câmbio: 1 euro em reais
- Informações sobre entregas de pacotes: 8167 2278 6764

Intenção a partir da digitação de consultas

- Cálculos: 5+4
- Conversão de unidades: 1 kg em pounds
- Conversão de câmbio: 1 euro em reais
- Informações sobre entregas de pacotes: 8167 2278 6764
- Informações sobre vôos: AZUL 454

Intenção a partir da digitação de consultas

- Cálculos: 5+4
- Conversão de unidades: 1 kg em pounds
- Conversão de câmbio: 1 euro em reais
- Informações sobre entregas de pacotes: 8167 2278 6764
- Informações sobre vôos: AZUL 454
- Código de área: 613

Intenção a partir da digitação de consultas

- Cálculos: 5+4
- Conversão de unidades: 1 kg em pounds
- Conversão de câmbio: 1 euro em reais
- Informações sobre entregas de pacotes: 8167 2278 6764
- Informações sobre vôos: AZUL 454
- Código de área: 613
- Mapa: uberlandia mg

Intenção a partir da digitação de consultas

- Cálculos: 5+4
- Conversão de unidades: 1 kg em pounds
- Conversão de câmbio: 1 euro em reais
- Informações sobre entregas de pacotes: 8167 2278 6764
- Informações sobre vôos: AZUL 454
- Código de área: 613
- Mapa: uberlandia mg
- Preço de ações: PETR3

Intenção a partir da digitação de consultas

- Cálculos: 5+4
- Conversão de unidades: 1 kg em pounds
- Conversão de câmbio: 1 euro em reais
- Informações sobre entregas de pacotes: 8167 2278 6764
- Informações sobre vôos: AZUL 454
- Código de área: 613
- Mapa: uberlandia mg
- Preço de ações: PETR3
- Filmes: Os estagiários

Como usar contexto para modificar os resultados?

- Restrição de resultados: não considerar os resultados inapropriados

Como usar contexto para modificar os resultados?

- Restrição de resultados: não considerar os resultados inapropriados
  - Para usuários no google.fr ...

Como usar contexto para modificar os resultados?

- Restrição de resultados: não considerar os resultados inapropriados
  - Para usuários no google.fr . . .
  - . . . somente mostrar resultados .fr

Como usar contexto para modificar os resultados?

- Restrição de resultados: não considerar os resultados inapropriados
  - Para usuários no google.fr . . .
  - . . . somente mostrar resultados .fr
- Modulação de ranking: usar um ranking genérico e reordenar baseado no contexto pessoal

Como usar contexto para modificar os resultados?

- Restrição de resultados: não considerar os resultados inapropriados
  - Para usuários no google.fr . . .
  - . . . somente mostrar resultados .fr
- Modulação de ranking: usar um ranking genérico e reordenar baseado no contexto pessoal
- Contextualização / personalização é uma área com muito potencial para melhorias/pesquisa

Conteúdo

Usuários da busca web

- Usam consultas curtas (média < 3)

Usuários da busca web

- Usam consultas curtas (média < 3)
- Raramente usam operadores

Usuários da busca web

- Usam consultas curtas (média < 3)
- Raramente usam operadores
- Não querem gastar tempo escrevendo consulta

Usuários da busca web

- Usam consultas curtas (média < 3)
- Raramente usam operadores
- Não querem gastar tempo escrevendo consulta
- Olham somente os primeiros resultados

Usuários da busca web

- Usam consultas curtas (média < 3)
- Raramente usam operadores
- Não querem gastar tempo escrevendo consulta
- Olham somente os primeiros resultados
- Querem uma interface limpa e não um buscador sobrecarregado

Usuários da busca web

- Usam consultas curtas (média < 3)
- Raramente usam operadores
- Não querem gastar tempo escrevendo consulta
- Olham somente os primeiros resultados
- Querem uma interface limpa e não um buscador sobrecarregado
- Variabilidade em termos de necessidades de usuários, expectivas, experiência, conhecimento ...

Usuários da busca web

- Usam consultas curtas (média < 3)
- Raramente usam operadores
- Não querem gastar tempo escrevendo consulta
- Olham somente os primeiros resultados
- Querem uma interface limpa e não um buscador sobrecarregado
- Variabilidade em termos de necessidades de usuários, expectivas, experiência, conhecimento . . .
  - 1º/3º mundo, Inglês/Egípcio, velho/jovem, rico/pobre, diferenças culturais

Usuários da busca web

- Usam consultas curtas (média < 3)
- Raramente usam operadores
- Não querem gastar tempo escrevendo consulta
- Olham somente os primeiros resultados
- Querem uma interface limpa e não um buscador sobrecarregado
- Variabilidade em termos de necessidades de usuários, expectivas, experiência, conhecimento . . .
  - 1º/3º mundo, Inglês/Egípcio, velho/jovem, rico/pobre, diferenças culturais
- Uma interface para necessidades divergentes

Como usuários avaliam buscadores?

- Relevância clássica em ORI (medida por F) pode também ser usado para a web

Como usuários avaliam buscadores?

- Relevância clássica em ORI (medida por F) pode também ser usado para a web
- Igualmente importante: confiabilidade, eliminação de conteúdo duplicado, conteúdo legível, carrega rápido, sem interferências visuais com pop-ups

Como usuários avaliam buscadores?

- Relevância clássica em ORI (medida por F) pode também ser usado para a web
- Igualmente importante: confiabilidade, eliminação de conteúdo duplicado, conteúdo legível, carrega rápido, sem interferências visuais com pop-ups
- Na web, taxa de precisão é mais importante que taxa de recuperação

Como usuários avaliam buscadores?

- Relevância clássica em ORI (medida por F) pode também ser usado para a web
- Igualmente importante: confiabilidade, eliminação de conteúdo duplicado, conteúdo legível, carrega rápido, sem interferências visuais com pop-ups
- Na web, taxa de precisão é mais importante que taxa de recuperação
  - Precisão em 1, precisão em 10, precisão nas 3 primeiras páginas de ranking

Necessidades de busca na web que requerem alta taxa de recuperação

Há casos em que recuperação é importante.

Exemplos:

Necessidades de busca na web que requerem alta taxa de recuperação

Há casos em que recuperação é importante.

Exemplos:

- Esta ideia foi patenteada?

Necessidades de busca na web que requerem alta taxa de recuperação

Há casos em que recuperação é importante.

Exemplos:

- Esta ideia foi patenteada?
- busca por consultores financeiros

Necessidades de busca na web que requerem alta taxa de recuperação

Há casos em que recuperação é importante.

Exemplos:

- Esta ideia foi patenteada?
- busca por consultores financeiros
- busca por informações sobre potenciais empregados

Necessidades de busca na web que requerem alta taxa de recuperação

Há casos em que recuperação é importante.

Exemplos:

- Esta ideia foi patenteada?
- busca por consultores financeiros
- busca por informações sobre potenciais empregados
- busca por informações sobre empresa para trabalhar

Conteúdo

Documentos Web

- Criação de conteúdo distribuída: sem coordenação ou planejamento

Documentos Web

- Criação de conteúdo distribuída: sem coordenação ou planejamento
  - "Democratização da publicação"

Documentos Web

- Criação de conteúdo distribuída: sem coordenação ou planejamento
  - "Democratização da publicação"
  - Resultado: heterogeneidade extrema de documentos na web

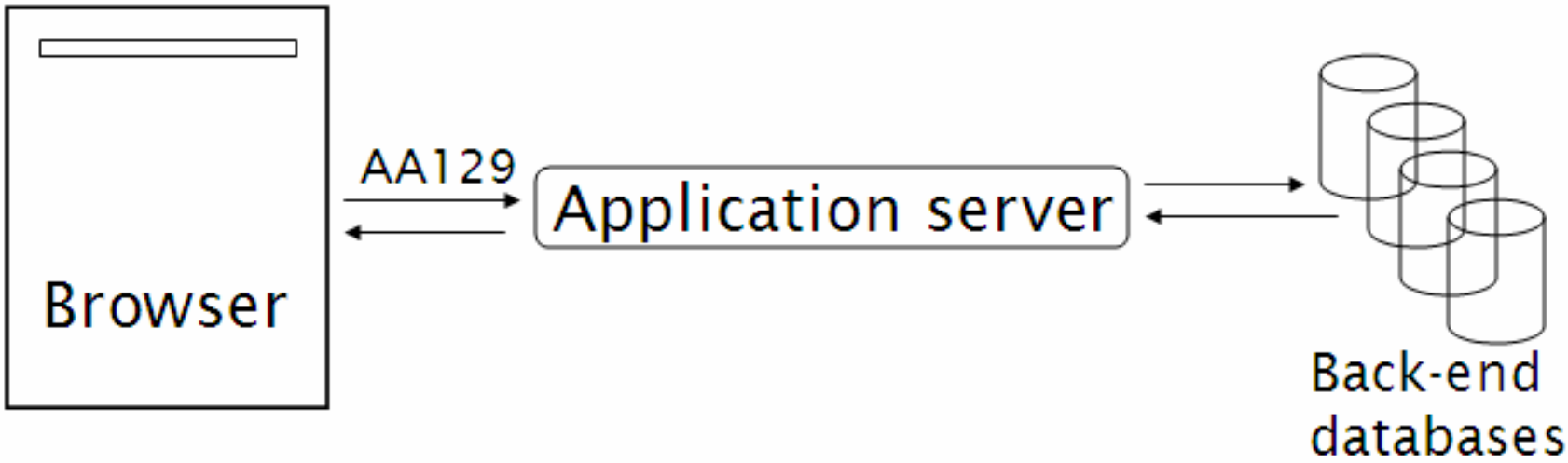
Documentos Web

- Criação de conteúdo distribuída: sem coordenação ou planejamento
  - "Democratização da publicação"
  - Resultado: heterogeneidade extrema de documentos na web
- Não estruturado (texto, html), semiestructurado (html, xml), estructurado/relacional (bancos de dados)

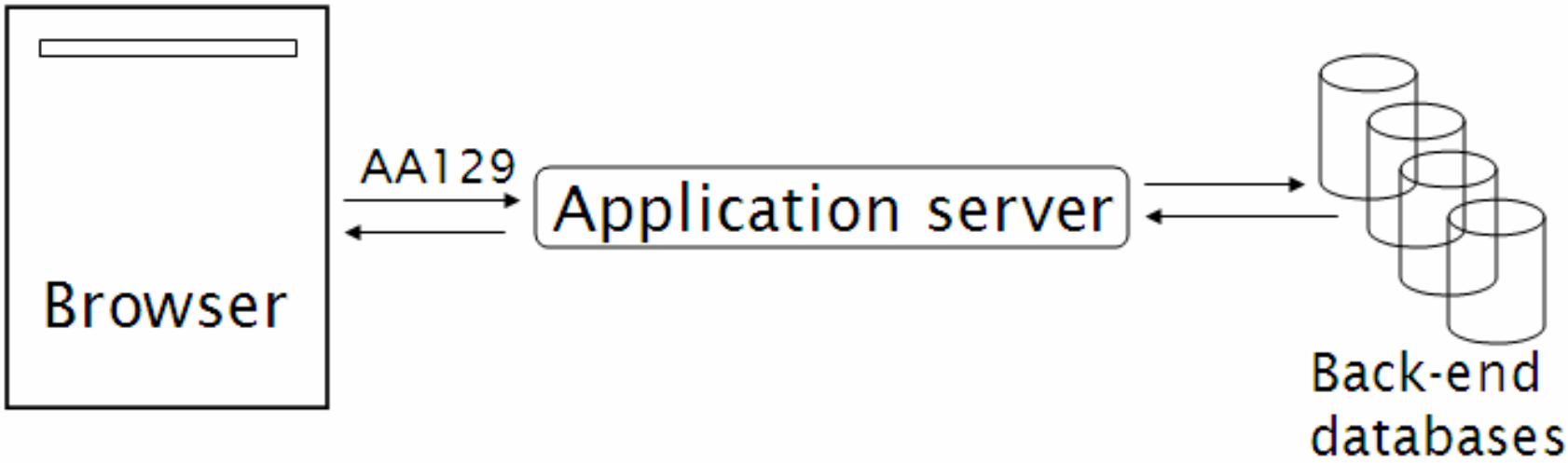
Documentos Web

- Criação de conteúdo distribuída: sem coordenação ou planejamento
  - "Democratização da publicação"
  - Resultado: heterogeneidade extrema de documentos na web
- Não estruturado (texto, html), semiestructurado (html, xml), estructurado/relacional (bancos de dados)
- Conteúdo dinamicamente gerado

Conteúdo dinâmico

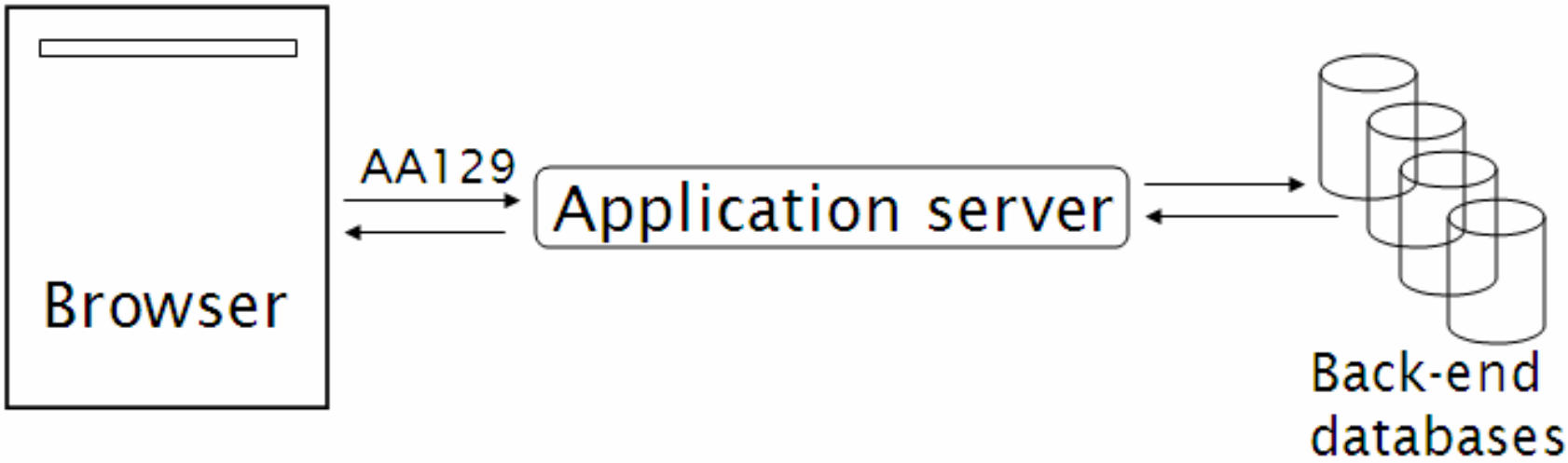


Conteúdo dinâmico



- Páginas dinâmicas são geradas quando o usuário pede – comumente a partir de um banco de dados

Conteúdo dinâmico



- Páginas dinâmicas são geradas quando o usuário pede – comumente a partir de um banco de dados
- Exemplo: estado atual do voo VRG 454

Conteúdo dinâmico (2)

- Maior parte do conteúdo dinâmico é ignorado pelos buscadores

Conteúdo dinâmico (2)

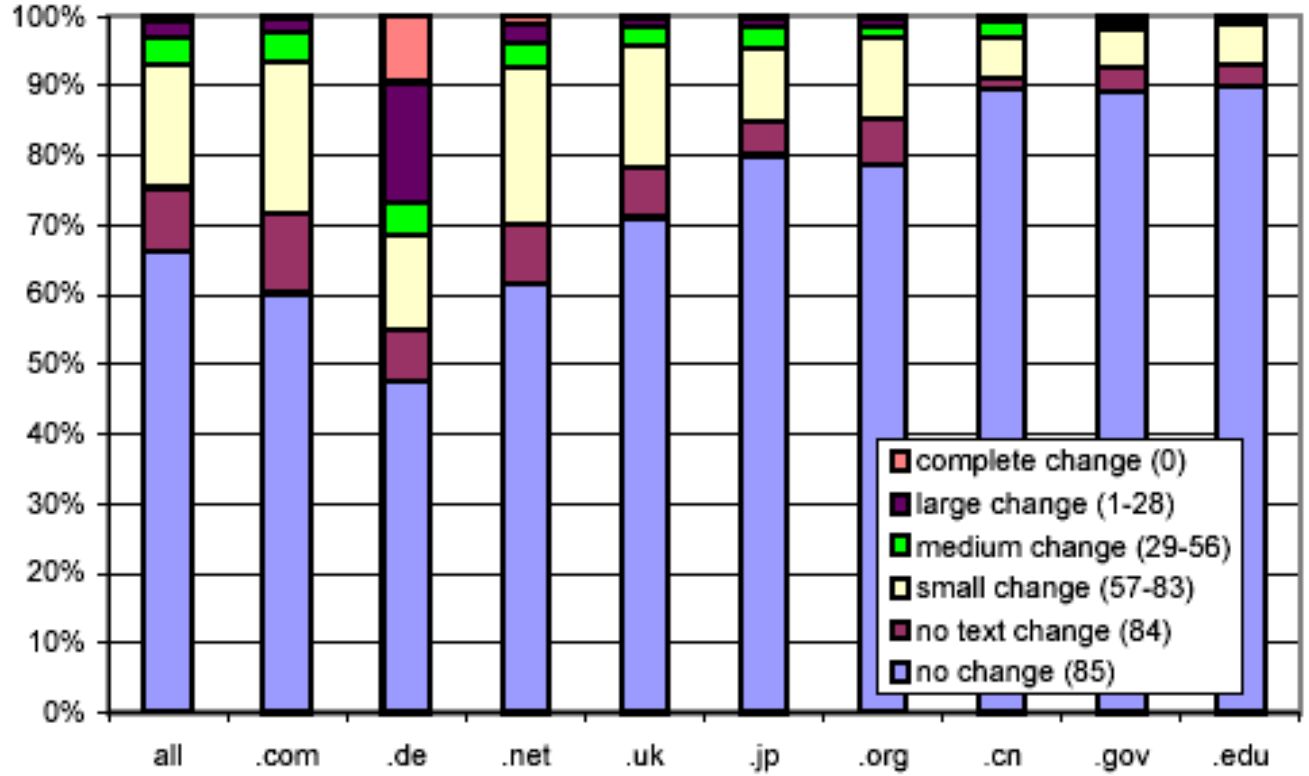
- Maior parte do conteúdo dinâmico é ignorado pelos buscadores
  - Muito para indexar

Conteúdo dinâmico (2)

- Maior parte do conteúdo dinâmico é ignorado pelos buscadores
  - Muito para indexar
- Na verdade, grande parte do conteúdo aparentemente estático também é gerado durante a requisição do usuário (asp, php etc.: cabeçalhos, data, anúncios etc)

Páginas Web pages mudam frequentemente

A Large-Scale Study of the Evolution of Web Pages, Fetterly 1997



Mudanças em 10 dias.

Pluri-idiomas

- Documentos em um grande número de idiomas

Pluri-idiomas

- Documentos em um grande número de idiomas
- Consultas em um grande número de idiomas

Pluri-idiomas

- Documentos em um grande número de idiomas
- Consultas em um grande número de idiomas
- Primeira abordagem: obter documentos do idioma usado na consulta

Pluri-idiomas

- Documentos em um grande número de idiomas
- Consultas em um grande número de idiomas
- Primeira abordagem: obter documentos do idioma usado na consulta
- Porém: frequentemente há conteúdo em idiomas diferentes do usado na consulta

Pluri-idiomas

- Documentos em um grande número de idiomas
- Consultas em um grande número de idiomas
- Primeira abordagem: obter documentos do idioma usado na consulta
- Porém: frequentemente há conteúdo em idiomas diferentes do usado na consulta
- Muitas pessoas conseguem entender, mas não pesquisar em um idioma diferente do nativo

Pluri-idiomas

- Documentos em um grande número de idiomas
- Consultas em um grande número de idiomas
- Primeira abordagem: obter documentos do idioma usado na consulta
- Porém: frequentemente há conteúdo em idiomas diferentes do usado na consulta
- Muitas pessoas conseguem entender, mas não pesquisar em um idioma diferente do nativo
- Tradução é importante

Duplicata de documentos

- Duplicação significativa – 30%–40% duplicatas em alguns estudos

Duplicata de documentos

- Duplicação significativa – 30%–40% duplicatas em alguns estudos
- Duplicatas na busca eram comuns no início da web

Duplicata de documentos

- Duplicação significativa – 30%–40% duplicatas em alguns estudos
- Duplicatas na busca eram comuns no início da web
- Hoje buscadores eliminam duplicatas bem

Duplicata de documentos

- Duplicação significativa – 30%–40% duplicatas em alguns estudos
- Duplicatas na busca eram comuns no início da web
- Hoje buscadores eliminam duplicatas bem
- Importante para satisfação dos usuários finais

Confiabilidade

- Para muitas coleções, é fácil avaliar a vericidade de um documento

Confiabilidade

- Para muitas coleções, é fácil avaliar a vericidade de um documento
  - Uma coleção de artigos da agência de notícias Reuters

Confiabilidade

- Para muitas coleções, é fácil avaliar a vericidade de um documento
  - Uma coleção de artigos da agência de notícias Reuters
  - Coleção de emails dos últimos três anos

Confiabilidade

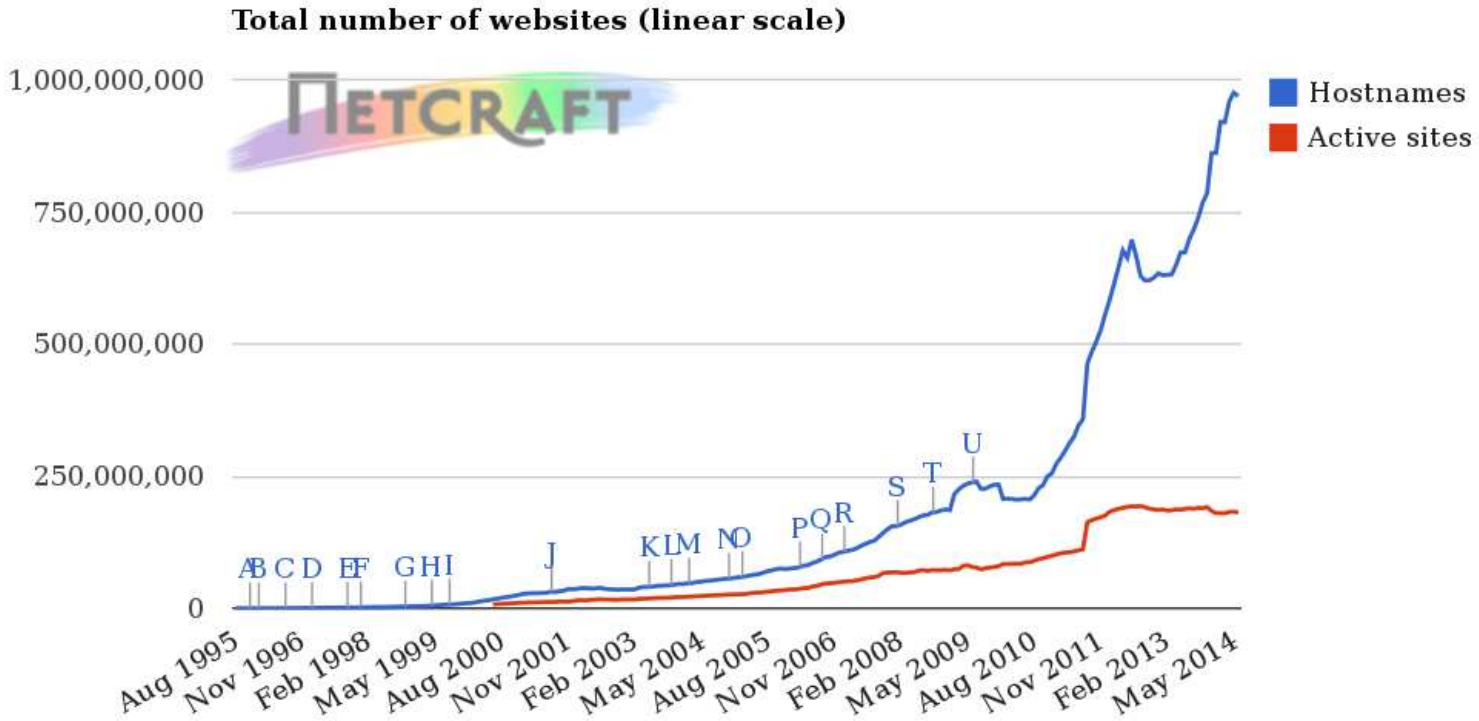
- Para muitas coleções, é fácil avaliar a vericidade de um documento
  - Uma coleção de artigos da agência de notícias Reuters
  - Coleção de emails dos últimos três anos
- Documentos da Web: em muitos casos, não sabemos como avaliar a informação

Confiabilidade

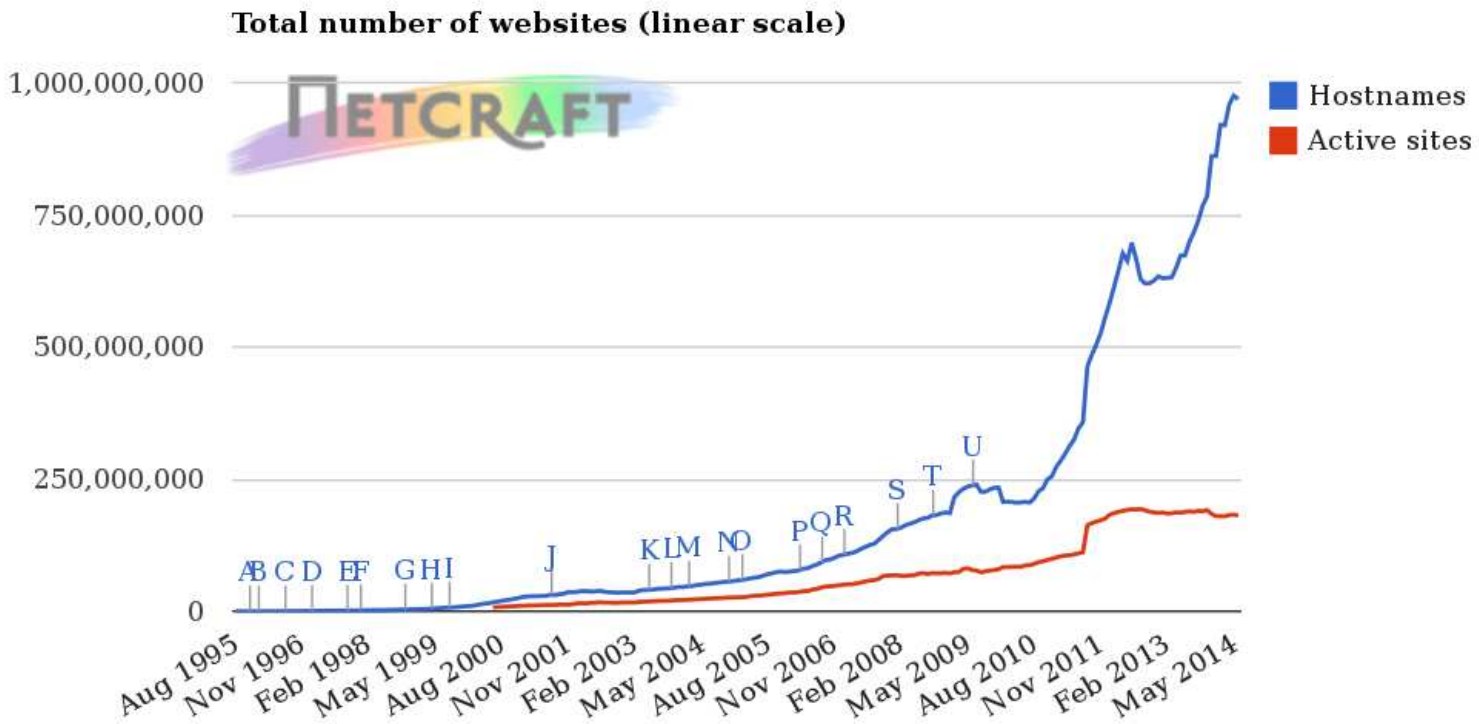
- Para muitas coleções, é fácil avaliar a vericidade de um documento
  - Uma coleção de artigos da agência de notícias Reuters
  - Coleção de emails dos últimos três anos
- Documentos da Web: em muitos casos, não sabemos como avaliar a informação
- Exemplo: boatos

Conteúdo

Crescimento da web

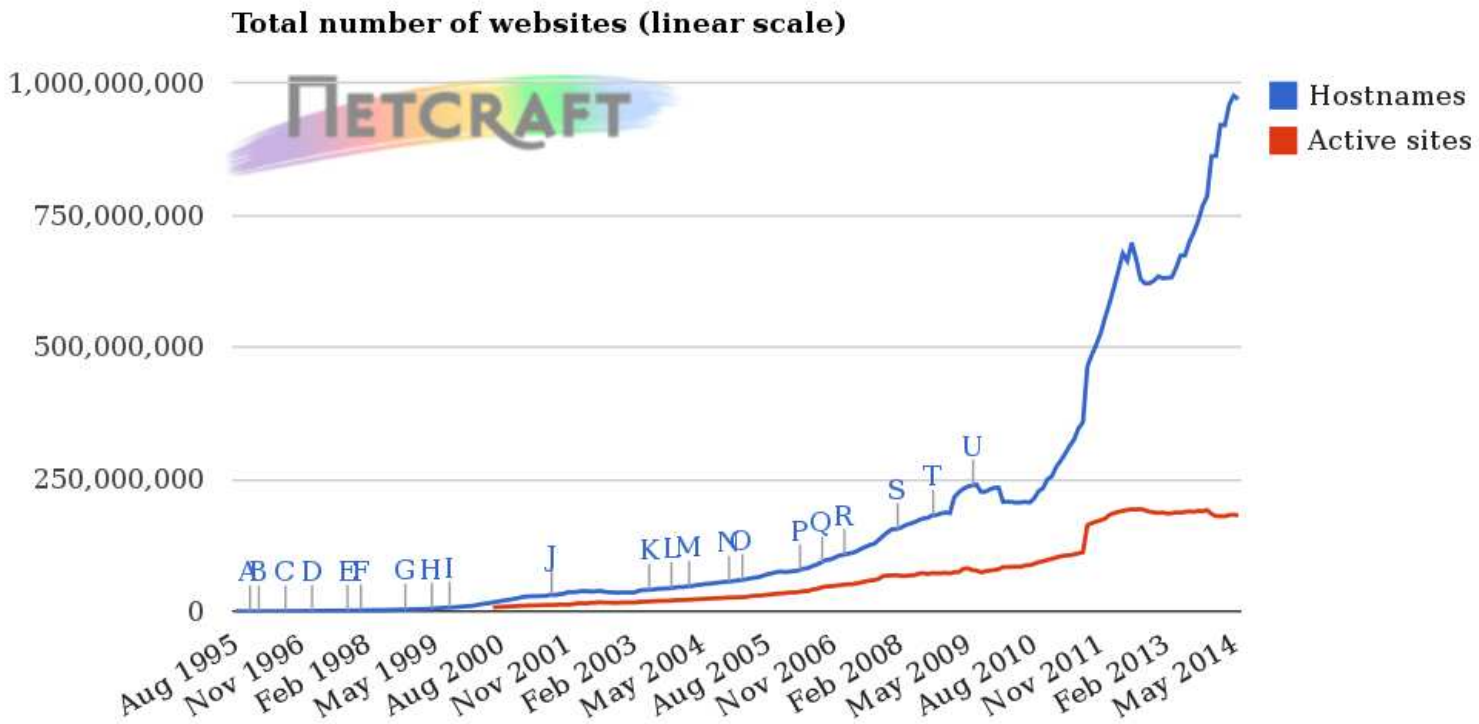


Crescimento da web



- A web continua crescendo.

Crescimento da web



- A web continua crescendo.
- Mas o crescimento é exponencial?

Tamanho da web: interessados

- Mídia

Tamanho da web: interessados

- Mídia
- Usuários

Tamanho da web: interessados

- Mídia
- Usuários
  - Público pode trocar de buscador para melhor cobertura da web

Tamanho da web: interessados

- Mídia
- Usuários
  - Público pode trocar de buscador para melhor cobertura da web
  - Usuários às vezes valorizam a taxa de recuperação. Se subestimarmos o tamanho da web, resultados do buscador podem ter baixa taxa de recuperação

Tamanho da web: interessados

- Mídia
- Usuários
  - Público pode trocar de buscador para melhor cobertura da web
  - Usuários às vezes valorizam a taxa de recuperação. Se subestimarmos o tamanho da web, resultados do buscador podem ter baixa taxa de recuperação
- Projetistas de buscadores – Quantas páginas necessito ser capaz de processar?

Tamanho da web: interessados

- Mídia
- Usuários
  - Público pode trocar de buscador para melhor cobertura da web
  - Usuários às vezes valorizam a taxa de recuperação. Se subestimarmos o tamanho da web, resultados do buscador podem ter baixa taxa de recuperação
- Projetistas de buscadores – Quantas páginas necessito ser capaz de processar?
- Projetistas de Crawler – rastreadores de páginas web (qual estratégia irá capturar aproximadamente essas N páginas?)

Tamanho da web

- O que determina o tamanho? Número de servidores web? Número de páginas? Terabytes de dados disponíveis?

Tamanho da web

- O que determina o tamanho? Número de servidores web? Número de páginas? Terabytes de dados disponíveis?
- Alguns servidores são raramente conectados

Tamanho da web

- O que determina o tamanho? Número de servidores web? Número de páginas? Terabytes de dados disponíveis?
- Alguns servidores são raramente conectados
  - Exemplo: Se o seu celular está rodando um servidor web, ele é parte da web?

Tamanho da web

- O que determina o tamanho? Número de servidores web? Número de páginas? Terabytes de dados disponíveis?
- Alguns servidores são raramente conectados
  - Exemplo: Se o seu celular está rodando um servidor web, ele é parte da web?
- A web dinâmica é infinita

Método simples para determinar um limite inferior

- Consulta OR de palavras frequentes em diversos idiomas

Método simples para determinar um limite inferior

- Consulta OR de palavras frequentes em diversos idiomas
  - Exemplo, se `the` está presente em cerca de 67.61% dos docs

Método simples para determinar um limite inferior

- Consulta OR de palavras frequentes em diversos idiomas
  - Exemplo, se `the` está presente em cerca de 67.61% dos docs
  - Então, se busca por `the` retornar 12 bilhões de páginas, temos cerca de 18 bilhões de documentos indexados

Método simples para determinar um limite inferior

- Consulta OR de palavras frequentes em diversos idiomas
  - Exemplo, se `the` está presente em cerca de 67.61% dos docs
  - Então, se busca por `the` retornar 12 bilhões de páginas, temos cerca de 18 bilhões de documentos indexados
- De acordo que esse tipo de consulta: tamanho da web $\geq$ 21,450,000,000 em 2007 em $\geq$ 25,350,000,000 em 2008 e $\geq 40,000,000,000$ em 2013

Método simples para determinar um limite inferior

- Consulta OR de palavras frequentes em diversos idiomas
  - Exemplo, se `the` está presente em cerca de 67.61% dos docs
  - Então, se busca por `the` retornar 12 bilhões de páginas, temos cerca de 18 bilhões de documentos indexados
- De acordo que esse tipo de consulta: tamanho da web $\geq$ 21,450,000,000 em 2007 em $\geq$ 25,350,000,000 em 2008 e $\geq 40,000,000,000$ em 2013
- Para contagem recentes e mais precisas: `http://www.worldwidewebsize.com/`

Resumo

- Visão geral da web
- Relação entre buscadores e a web
- Anúncios
- Duplicatas e quase-duplicata
- Spam
- Perfil de consultas na web
- Estrutura e usuários da web
- Dinamicidade da web
- Crescimento e tamanho da web