

# Introduction to Information Retrieval

<http://informationretrieval.org>

## IIR 19: Web Search Basics

Hinrich Schütze

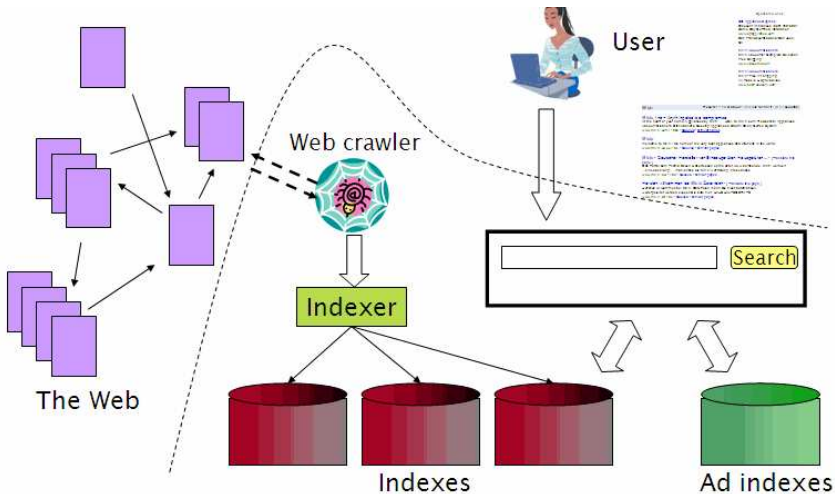
Institute for Natural Language Processing, Universität Stuttgart

2008.07.07

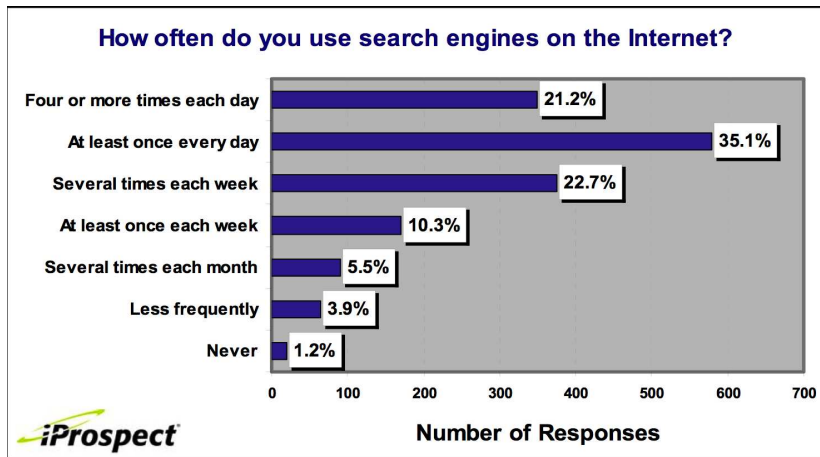
# Overview

- 1 Web IR
  - Links
  - Queries
  - Context
  - Users
  - Documents
  - Size
- 2 Ads & Spam
  - Ads
  - Spam

# Web search overview



# Search is a top activity on the web



# Without search engines, the web wouldn't work

- Without search, content is hard to find.

# Without search engines, the web wouldn't work

- Without search, content is hard to find.
- → Without search, there is no incentive to create content.

# Without search engines, the web wouldn't work

- Without search, content is hard to find.
- → Without search, there is no incentive to create content.
  - Why publish something if nobody will read it?

# Without search engines, the web wouldn't work

- Without search, content is hard to find.
- → Without search, there is no incentive to create content.
  - Why publish something if nobody will read it?
  - Why publish something if I don't get ad revenue from it?



# Without search engines, the web wouldn't work

- Without search, content is hard to find.
- → Without search, there is no incentive to create content.
  - Why publish something if nobody will read it?
  - Why publish something if I don't get ad revenue from it?
- Interest aggregation

# Without search engines, the web wouldn't work

- Without search, content is hard to find.
- → Without search, there is no incentive to create content.
  - Why publish something if nobody will read it?
  - Why publish something if I don't get ad revenue from it?
- Interest aggregation
  - Unique feature of the web: A small number of geographically dispersed people with similar interests can find each other.

# Without search engines, the web wouldn't work

- Without search, content is hard to find.
- → Without search, there is no incentive to create content.
  - Why publish something if nobody will read it?
  - Why publish something if I don't get ad revenue from it?
- Interest aggregation
  - Unique feature of the web: A small number of geographically dispersed people with similar interests can find each other.
  - Elementary school kids with hemophilia

# Without search engines, the web wouldn't work

- Without search, content is hard to find.
- → Without search, there is no incentive to create content.
  - Why publish something if nobody will read it?
  - Why publish something if I don't get ad revenue from it?
- Interest aggregation
  - Unique feature of the web: A small number of geographically dispersed people with similar interests can find each other.
  - Elementary school kids with hemophilia
  - People interested in translating R5R5 Scheme into relatively portable C (open source project)

# Without search engines, the web wouldn't work

- Without search, content is hard to find.
- → Without search, there is no incentive to create content.
  - Why publish something if nobody will read it?
  - Why publish something if I don't get ad revenue from it?
- Interest aggregation
  - Unique feature of the web: A small number of geographically dispersed people with similar interests can find each other.
  - Elementary school kids with hemophilia
  - People interested in translating R5R5 Scheme into relatively portable C (open source project)
  - Interest aggregation without search engines is not possible.

# Without search engines, the web wouldn't work

- Without search, content is hard to find.
- → Without search, there is no incentive to create content.
  - Why publish something if nobody will read it?
  - Why publish something if I don't get ad revenue from it?
- Interest aggregation
  - Unique feature of the web: A small number of geographically dispersed people with similar interests can find each other.
  - Elementary school kids with hemophilia
  - People interested in translating R5R5 Scheme into relatively portable C (open source project)
  - Interest aggregation without search engines is not possible.
- Somebody needs to pay for the web.

# Without search engines, the web wouldn't work

- Without search, content is hard to find.
- → Without search, there is no incentive to create content.
  - Why publish something if nobody will read it?
  - Why publish something if I don't get ad revenue from it?
- Interest aggregation
  - Unique feature of the web: A small number of geographically dispersed people with similar interests can find each other.
  - Elementary school kids with hemophilia
  - People interested in translating R5R5 Scheme into relatively portable C (open source project)
  - Interest aggregation without search engines is not possible.
- Somebody needs to pay for the web.
  - Servers, web infrastructure, content creation

# Without search engines, the web wouldn't work

- Without search, content is hard to find.
- → Without search, there is no incentive to create content.
  - Why publish something if nobody will read it?
  - Why publish something if I don't get ad revenue from it?
- Interest aggregation
  - Unique feature of the web: A small number of geographically dispersed people with similar interests can find each other.
  - Elementary school kids with hemophilia
  - People interested in translating R5R5 Scheme into relatively portable C (open source project)
  - Interest aggregation without search engines is not possible.
- Somebody needs to pay for the web.
  - Servers, web infrastructure, content creation
  - A large part today is paid by search ads.



# Outline

- 1 Web IR
  - Links
  - Queries
  - Context
  - Users
  - Documents
  - Size
- 2 Ads & Spam
  - Ads
  - Spam

# Web IR: Differences from traditional IR

- Links: The web is a hyperlinked document collection.

# Web IR: Differences from traditional IR

- Links: The web is a hyperlinked document collection.
- Queries: Web queries are different, more varied and there are a lot of them.

# Web IR: Differences from traditional IR

- Links: The web is a hyperlinked document collection.
- Queries: Web queries are different, more varied and there are a lot of them.
- Users: Users are different, more varied and there are a lot of them.

# Web IR: Differences from traditional IR

- Links: The web is a hyperlinked document collection.
- Queries: Web queries are different, more varied and there are a lot of them.
- Users: Users are different, more varied and there are a lot of them.
- Documents: Documents are different, more varied and there are a lot of them.

# Web IR: Differences from traditional IR

- Links: The web is a hyperlinked document collection.
- Queries: Web queries are different, more varied and there are a lot of them.
- Users: Users are different, more varied and there are a lot of them.
- Documents: Documents are different, more varied and there are a lot of them.
- Context: Context is more important on the web than in many other IR applications.

# Web IR: Differences from traditional IR

- Links: The web is a hyperlinked document collection.
- Queries: Web queries are different, more varied and there are a lot of them.
- Users: Users are different, more varied and there are a lot of them.
- Documents: Documents are different, more varied and there are a lot of them.
- Context: Context is more important on the web than in many other IR applications.
- Ads and spam

# Web IR: Differences from traditional IR

- Links: The web is a hyperlinked document collection.
- Queries: Web queries are different, more varied and there are a lot of them. **How many?**
- Users: Users are different, more varied and there are a lot of them. **How many?**
- Documents: Documents are different, more varied and there are a lot of them. **How many?**
- Context: Context is more important on the web than in many other IR applications.
- Ads and spam



# Web IR: Differences from traditional IR

- Links: The web is a hyperlinked document collection.
- Queries: Web queries are different, more varied and there are a lot of them. How many?  $10^8$  every day, approaching  $10^9$
- Users: Users are different, more varied and there are a lot of them. How many?  $10^9$
- Documents: Documents are different, more varied and there are a lot of them. How many?  $\approx 10^{11}$ . Indexed:  $10^{10}$
- Context: Context is more important on the web than in many other IR applications.
- Ads and spam

# Outline

- 1 Web IR
  - Links
  - Queries
  - Context
  - Users
  - Documents
  - Size
- 2 Ads & Spam
  - Ads
  - Spam

# Search in a hyperlinked collection

- Web search in most cases is interleaved with navigation . . .

# Search in a hyperlinked collection

- Web search in most cases is interleaved with navigation ...
- ... i.e., with following links.

# Search in a hyperlinked collection

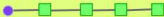
- Web search in most cases is interleaved with navigation ...
- ... i.e., with following links.
- Different from most other IR collections

# Kinds of behaviors we see in the data

Short / Nav



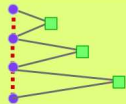
Topic exploration



Topic switch



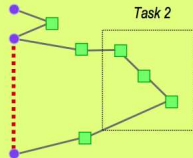
Methodical results exploration



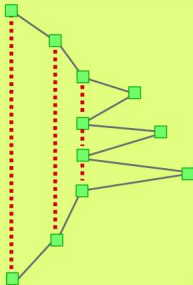
Query reform



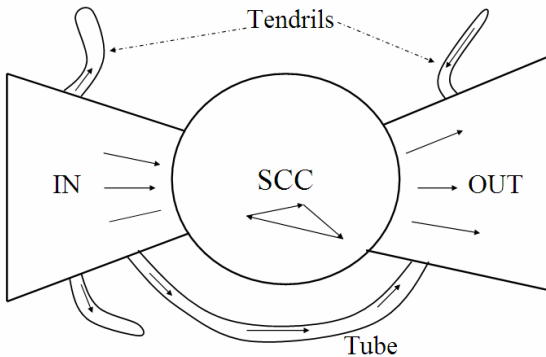
Multitasking



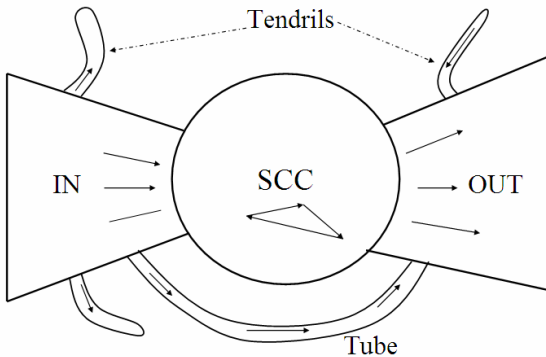
Stacking behavior



# Bowtie structure of the web



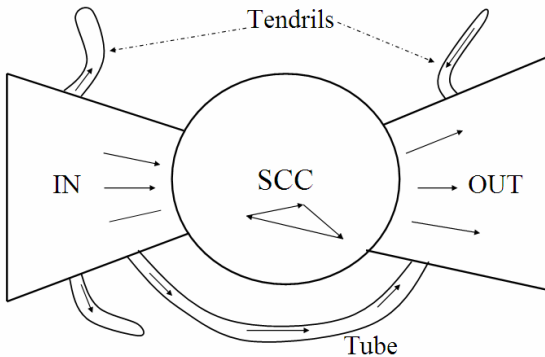
# Bowtie structure of the web



- Strongly connected component (SCC) in the center

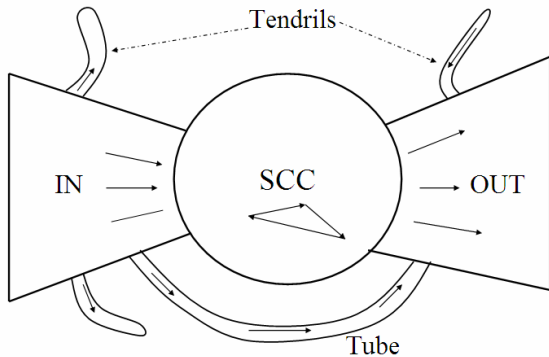


# Bowtie structure of the web



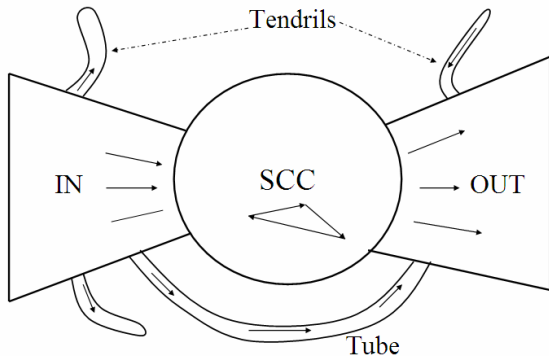
- Strongly connected component (SCC) in the center
- Lots of pages that get linked to, but don't link (OUT)

# Bowtie structure of the web



- Strongly connected component (SCC) in the center
- Lots of pages that get linked to, but don't link (OUT)
- Lots of pages that link to other pages, but don't get linked to (IN)

# Bowtie structure of the web



- Strongly connected component (SCC) in the center
- Lots of pages that get linked to, but don't link (OUT)
- Lots of pages that link to other pages, but don't get linked to (IN)
- Tendrils, tubes, islands

# Outline

- 1 Web IR
  - Links
  - Queries
  - Context
  - Users
  - Documents
  - Size
- 2 Ads & Spam
  - Ads
  - Spam

# Query distribution (1)

Most frequent queries on a large search engine on 2002.10.26.

1	sex	16	crack	31	juegos	46	Caramail
2	(artifact)	17	games	32	nude	47	msn
3	(artifact)	18	pussy	33	music	48	jennifer lopez
4	porno	19	cracks	34	musica	49	tits
5	mp3	20	lolita	35	anal	50	free porn
6	Halloween	21	britney spears	36	free6	51	cheats
7	sexo	22	ebay	37	avril lavigne	52	yahoo.com
8	chat	23	sexe	38	www.hotmail.com	53	eminem
9	porn	24	Pamela Anderson	39	winzip	54	Christina Aguilera
10	yahoo	25	warez	40	fuck	55	incest
11	KaZaA	26	divx	41	wallpaper	56	letras de canciones
12	xxx	27	gay	42	hotmail.com	57	hardcore
13	Hentai	28	harry potter	43	postales	58	weather
14	lyrics	29	playboy	44	shakira	59	wallpapers
15	hotmail	30	lolitas	45	traductor	60	lingerie

# Query distribution (1)

Most frequent queries on a large search engine on 2002.10.26.

1	sex	16	crack	31	juegos	46	Caramail
2	(artifact)	17	games	32	nude	47	msn
3	(artifact)	18	pussy	33	music	48	jennifer lopez
4	porno	19	cracks	34	musica	49	tits
5	mp3	20	lolita	35	anal	50	free porn
6	Halloween	21	britney spears	36	free6	51	cheats
7	sexo	22	ebay	37	avril lavigne	52	yahoo.com
8	chat	23	sexe	38	www.hotmail.com	53	eminem
9	porn	24	Pamela Anderson	39	winzip	54	Christina Aguilera
10	yahoo	25	warez	40	fuck	55	incest
11	KaZaA	26	divx	41	wallpaper	56	letras de canciones
12	xxx	27	gay	42	hotmail.com	57	hardcore
13	Hentai	28	harry potter	43	postales	58	weather
14	lyrics	29	playboy	44	shakira	59	wallpapers
15	hotmail	30	lolitas	45	traductor	60	lingerie

More than 1/3 of these are queries for adult content.

# Query distribution (1)

Most frequent queries on a large search engine on 2002.10.26.

1	sex	16	crack	31	juegos	46	Caramail
2	(artifact)	17	games	32	nude	47	msn
3	(artifact)	18	pussy	33	music	48	jennifer lopez
4	porno	19	cracks	34	musica	49	tits
5	mp3	20	lolita	35	anal	50	free porn
6	Halloween	21	britney spears	36	free6	51	cheats
7	sexo	22	ebay	37	avril lavigne	52	yahoo.com
8	chat	23	sexe	38	www.hotmail.com	53	eminem
9	porn	24	Pamela Anderson	39	winzip	54	Christina Aguilera
10	yahoo	25	warez	40	fuck	55	incest
11	KaZaA	26	divx	41	wallpaper	56	letras de canciones
12	xxx	27	gay	42	hotmail.com	57	hardcore
13	Hentai	28	harry potter	43	postales	58	weather
14	lyrics	29	playboy	44	shakira	59	wallpapers
15	hotmail	30	lolitas	45	traductor	60	lingerie

More than 1/3 of these are queries for adult content. Does this mean that most people are looking for adult content?

## Query distribution (2)

- Queries have a power law distribution.



## Query distribution (2)

- Queries have a power law distribution.
- Recall Zipf's law: a few very frequent words, a large number of very rare words

## Query distribution (2)

- Queries have a power law distribution.
- Recall Zipf's law: a few very frequent words, a large number of very rare words
- Same here: a few very frequent queries, a large number of very rare queries

## Query distribution (2)

- Queries have a power law distribution.
- Recall Zipf's law: a few very frequent words, a large number of very rare words
- Same here: a few very frequent queries, a large number of very rare queries
- Examples of rare queries: search for names, towns, books etc

## Query distribution (2)

- Queries have a power law distribution.
- Recall Zipf's law: a few very frequent words, a large number of very rare words
- Same here: a few very frequent queries, a large number of very rare queries
- Examples of rare queries: search for names, towns, books etc
- The proportion of adult queries is much lower than  $1/3$

# Types of queries / user needs in web search

- **Informational user needs:** I need information on something.  
“low hemoglobin”

# Types of queries / user needs in web search

- **Informational user needs:** I need information on something.  
“low hemoglobin”
- We called this “information need” earlier in the class.

# Types of queries / user needs in web search

- **Informational user needs:** I need information on something.  
“low hemoglobin”
- We called this “information need” earlier in the class.
- **On the web, information needs proper are only a subclass of user needs.**

# Types of queries / user needs in web search

- **Informational user needs:** I need information on something.  
“low hemoglobin”
- We called this “information need” earlier in the class.
- **On the web, information needs proper are only a subclass of user needs.**
- Other user needs: Navigational and transactional



# Types of queries / user needs in web search

- **Informational user needs:** I need information on something.  
“low hemoglobin”
- We called this “information need” earlier in the class.
- **On the web, information needs proper are only a subclass of user needs.**
- Other user needs: Navigational and transactional
- **Navigational user needs:** I want to go to this web site.  
“hotmail”, “myspace”, “United Airlines”

# Types of queries / user needs in web search

- **Informational user needs:** I need information on something.  
“low hemoglobin”
- We called this “information need” earlier in the class.
- **On the web, information needs proper are only a subclass of user needs.**
- Other user needs: Navigational and transactional
- **Navigational user needs:** I want to go to this web site.  
“hotmail”, “myspace”, “United Airlines”
- **Transactional user needs:** I want to make a transaction.

# Types of queries / user needs in web search

- **Informational user needs:** I need information on something.  
“low hemoglobin”
- We called this “information need” earlier in the class.
- **On the web, information needs proper are only a subclass of user needs.**
- Other user needs: Navigational and transactional
- **Navigational user needs:** I want to go to this web site.  
“hotmail”, “myspace”, “United Airlines”
- **Transactional user needs:** I want to make a transaction.
  - Buy something: “MacBook Air”

# Types of queries / user needs in web search

- **Informational user needs:** I need information on something.  
“low hemoglobin”
- We called this “information need” earlier in the class.
- **On the web, information needs proper are only a subclass of user needs.**
- Other user needs: Navigational and transactional
- **Navigational user needs:** I want to go to this web site.  
“hotmail”, “myspace”, “United Airlines”
- **Transactional user needs:** I want to make a transaction.
  - Buy something: “MacBook Air”
  - Download something: “Acrobat Reader”

# Types of queries / user needs in web search

- **Informational user needs:** I need information on something.  
“low hemoglobin”
- We called this “information need” earlier in the class.
- **On the web, information needs proper are only a subclass of user needs.**
- Other user needs: Navigational and transactional
- **Navigational user needs:** I want to go to this web site.  
“hotmail”, “myspace”, “United Airlines”
- **Transactional user needs:** I want to make a transaction.
  - Buy something: “MacBook Air”
  - Download something: “Acrobat Reader”
  - Chat with someone: “live soccer chat”

# Types of queries / user needs in web search

- **Informational user needs:** I need information on something.  
“low hemoglobin”
- We called this “information need” earlier in the class.
- **On the web, information needs proper are only a subclass of user needs.**
- Other user needs: Navigational and transactional
- **Navigational user needs:** I want to go to this web site.  
“hotmail”, “myspace”, “United Airlines”
- **Transactional user needs:** I want to make a transaction.
  - Buy something: “MacBook Air”
  - Download something: “Acrobat Reader”
  - Chat with someone: “live soccer chat”
- **Difficult problem:** How can the search engine tell what the user need or intent for a particular query is?

# Outline

- 1 Web IR
  - Links
  - Queries
  - **Context**
  - Users
  - Documents
  - Size
  
- 2 Ads & Spam
  - Ads
  - Spam

# User intent: Answering the need behind the query

- What can we do to guess user intent?



# User intent: Answering the need behind the query

- What can we do to guess user intent?
- Guess user intent independent of context:

# User intent: Answering the need behind the query

- What can we do to guess user intent?
- Guess user intent independent of context:
  - Spell correction

# User intent: Answering the need behind the query

- What can we do to guess user intent?
- Guess user intent independent of context:
  - Spell correction
  - Precomputed “typing” of queries (next slide)

# User intent: Answering the need behind the query

- What can we do to guess user intent?
- Guess user intent independent of context:
  - Spell correction
  - Precomputed “typing” of queries (next slide)
- Better: Guess user intent based on context:

# User intent: Answering the need behind the query

- What can we do to guess user intent?
- Guess user intent independent of context:
  - Spell correction
  - Precomputed “typing” of queries (next slide)
- Better: Guess user intent based on context:
  - Geographic context (slide after next)

# User intent: Answering the need behind the query

- What can we do to guess user intent?
- Guess user intent independent of context:
  - Spell correction
  - Precomputed “typing” of queries (next slide)
- Better: Guess user intent based on context:
  - Geographic context (slide after next)
  - Context of user in this session (e.g., previous query)

# User intent: Answering the need behind the query

- What can we do to guess user intent?
- Guess user intent independent of context:
  - Spell correction
  - Precomputed “typing” of queries (next slide)
- Better: Guess user intent based on context:
  - Geographic context (slide after next)
  - Context of user in this session (e.g., previous query)
  - Context provided by personal profile (Yahoo/MSN do this, Google claims it doesn't)

# Guessing of user intent by “typing” queries

- Calculation: 5+4
- Unit conversion: 1 kg in pounds
- Currency conversion: 1 euro in kronor
- Tracking number: 8167 2278 6764
- Flight info: LH 454
- Area code: 650
- Map: columbus oh
- Stock price: msft
- Albums/movies etc: coldplay



# The spatial context: Geo-search

- Three relevant locations

# The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)

# The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)

# The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)
  - User (located in Palo Alto)

# The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)
  - User (located in Palo Alto)
- Locating the user

# The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)
  - User (located in Palo Alto)
- Locating the user
  - IP address

# The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)
  - User (located in Palo Alto)
- Locating the user
  - IP address
  - Information provided by user (e.g., in user profile)

# The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)
  - User (located in Palo Alto)
- Locating the user
  - IP address
  - Information provided by user (e.g., in user profile)
  - Mobile phone



# The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)
  - User (located in Palo Alto)
- Locating the user
  - IP address
  - Information provided by user (e.g., in user profile)
  - Mobile phone
- **Geo-tagging**: Parse text and identify the coordinates of the geographic entities

# The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)
  - User (located in Palo Alto)
- Locating the user
  - IP address
  - Information provided by user (e.g., in user profile)
  - Mobile phone
- **Geo-tagging**: Parse text and identify the coordinates of the geographic entities
  - Example: East Palo Alto CA → Latitude: 37.47 N, Longitude: 122.14 W

# The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)
  - User (located in Palo Alto)
- Locating the user
  - IP address
  - Information provided by user (e.g., in user profile)
  - Mobile phone
- **Geo-tagging**: Parse text and identify the coordinates of the geographic entities
  - Example: East Palo Alto CA → Latitude: 37.47 N, Longitude: 122.14 W
  - Important NLP problem

# How do we use context to modify query results?

- Result restriction: Don't consider inappropriate results

# How do we use context to modify query results?

- Result restriction: Don't consider inappropriate results
  - For user on google.fr ...

# How do we use context to modify query results?

- Result restriction: Don't consider inappropriate results
  - For user on google.fr ...
  - ... only show .fr results

# How do we use context to modify query results?

- Result restriction: Don't consider inappropriate results
  - For user on google.fr ...
  - ... only show .fr results
- Ranking modulation: use a rough generic ranking, rerank based on personal context

# How do we use context to modify query results?

- Result restriction: Don't consider inappropriate results
  - For user on google.fr ...
  - ... only show .fr results
- Ranking modulation: use a rough generic ranking, rerank based on personal context
- Contextualization / personalization is an area of search with a lot of potential for improvement.



# Outline

- 1 Web IR
  - Links
  - Queries
  - Context
  - **Users**
  - Documents
  - Size
- 2 Ads & Spam
  - Ads
  - Spam

# Users of web search

- Use short queries (average  $< 3$ )

# Users of web search

- Use short queries (average  $< 3$ )
- Rarely use operators

# Users of web search

- Use short queries (average  $< 3$ )
- Rarely use operators
- Don't want to spend a lot of time on composing a query

# Users of web search

- Use short queries (average  $< 3$ )
- Rarely use operators
- Don't want to spend a lot of time on composing a query
- Only look at the first couple of results

# Users of web search

- Use short queries (average  $< 3$ )
- Rarely use operators
- Don't want to spend a lot of time on composing a query
- Only look at the first couple of results
- Want a simple UI, not a search engine start page overloaded with graphics

# Users of web search

- Use short queries (average  $< 3$ )
- Rarely use operators
- Don't want to spend a lot of time on composing a query
- Only look at the first couple of results
- Want a simple UI, not a search engine start page overloaded with graphics
- Extreme variability in terms of user needs, user expectations, experience, knowledge, . . .

# Users of web search

- Use short queries (average  $< 3$ )
- Rarely use operators
- Don't want to spend a lot of time on composing a query
- Only look at the first couple of results
- Want a simple UI, not a search engine start page overloaded with graphics
- Extreme variability in terms of user needs, user expectations, experience, knowledge, . . .
  - Industrial/developing world, English/Estonian, old/young, rich/poor, differences in culture and class



# Users of web search

- Use short queries (average  $< 3$ )
- Rarely use operators
- Don't want to spend a lot of time on composing a query
- Only look at the first couple of results
- Want a simple UI, not a search engine start page overloaded with graphics
- Extreme variability in terms of user needs, user expectations, experience, knowledge, . . .
  - Industrial/developing world, English/Estonian, old/young, rich/poor, differences in culture and class
- One interface for hugely divergent needs

# How do users evaluate search engines?

- Classic IR relevance (as measured by  $F$ ) can also be used for web IR.

# How do users evaluate search engines?

- Classic IR relevance (as measured by  $F$ ) can also be used for web IR.
- Equally important: Trust, duplicate elimination, readability, loads fast, no pop-ups

# How do users evaluate search engines?

- Classic IR relevance (as measured by  $F$ ) can also be used for web IR.
- Equally important: Trust, duplicate elimination, readability, loads fast, no pop-ups
- On the web, precision is more important than recall.

# How do users evaluate search engines?

- Classic IR relevance (as measured by  $F$ ) can also be used for web IR.
- Equally important: Trust, duplicate elimination, readability, loads fast, no pop-ups
- On the web, precision is more important than recall.
  - Precision at 1, precision at 10, precision on the first 2-3 pages

# How do users evaluate search engines?

- Classic IR relevance (as measured by  $F$ ) can also be used for web IR.
- Equally important: Trust, duplicate elimination, readability, loads fast, no pop-ups
- On the web, precision is more important than recall.
  - Precision at 1, precision at 10, precision on the first 2-3 pages
  - But there is a subset of queries where recall matters.

# Web information needs that require high recall?



# Web information needs that require high recall?

- ?
- Has this idea been patented?



# Web information needs that require high recall?

- ?
- Has this idea been patented?
- Searching for info on a prospective financial advisor

# Web information needs that require high recall?

- ?
- Has this idea been patented?
- Searching for info on a prospective financial advisor
- Searching for info on a prospective employee

# Web information needs that require high recall?

- ?
- Has this idea been patented?
- Searching for info on a prospective financial advisor
- Searching for info on a prospective employee
- Searching for info on a date

# Outline

- 1 Web IR
  - Links
  - Queries
  - Context
  - Users
  - Documents
  - Size
  
- 2 Ads & Spam
  - Ads
  - Spam

# Web documents: different from other IR collections

- Distributed content creation: no design, no co-ordination

# Web documents: different from other IR collections

- Distributed content creation: no design, no co-ordination
  - “Democratization of publishing”

# Web documents: different from other IR collections

- Distributed content creation: no design, no co-ordination
  - “Democratization of publishing”
  - Result: extreme heterogeneity of documents on the web

# Web documents: different from other IR collections

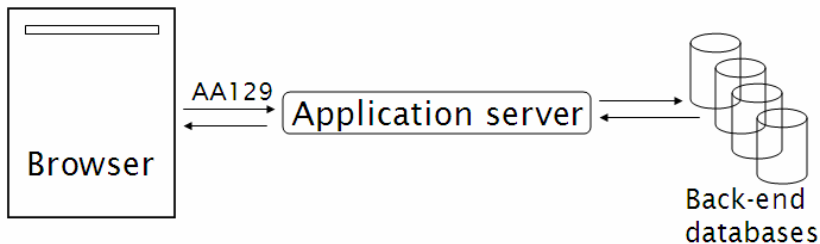
- Distributed content creation: no design, no co-ordination
  - “Democratization of publishing”
  - Result: extreme heterogeneity of documents on the web
- Unstructured (text, html), semistructured (html, xml), structured/relational (databases)



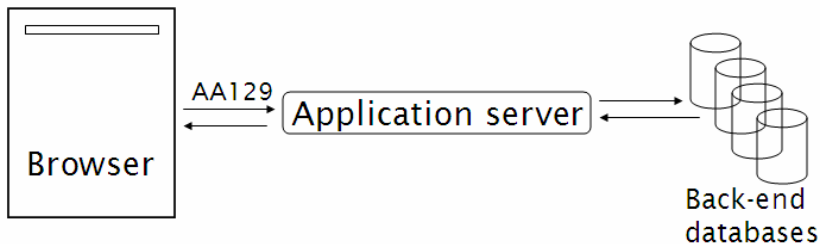
# Web documents: different from other IR collections

- Distributed content creation: no design, no co-ordination
  - “Democratization of publishing”
  - Result: extreme heterogeneity of documents on the web
- Unstructured (text, html), semistructured (html, xml), structured/relational (databases)
- Dynamically generated content

# Dynamic content

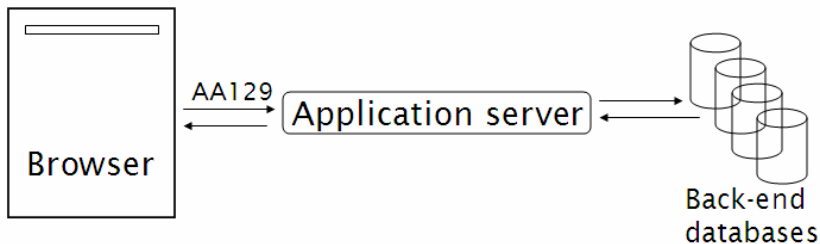


# Dynamic content



- Dynamic pages are generated from scratch when the user requests them – usually from underlying data in a database.

# Dynamic content



- Dynamic pages are generated from scratch when the user requests them – usually from underlying data in a database.
- Example: current status of flight LH 454

## Dynamic content (2)

- Most (truly) dynamic content is ignored by web spiders.

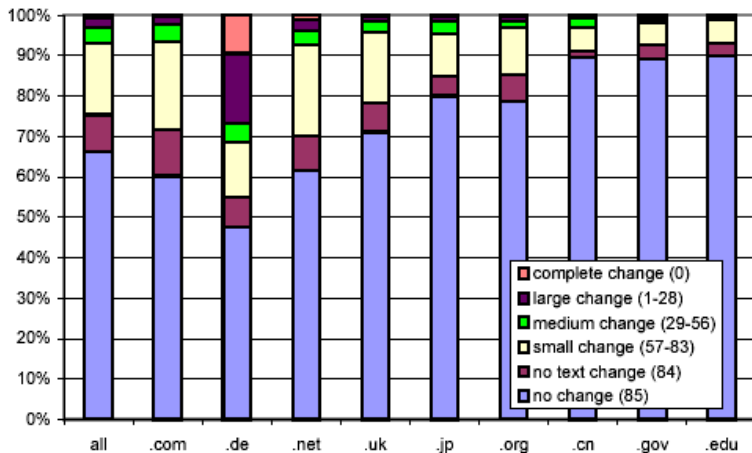
## Dynamic content (2)

- Most (truly) dynamic content is ignored by web spiders.
  - It's too much to index it all.

## Dynamic content (2)

- Most (truly) dynamic content is ignored by web spiders.
  - It's too much to index it all.
- Actually, a lot of “static” content is also assembled on the fly (asp, php etc.: headers, date, ads etc)

# Web pages change frequently (Fetterly 1997)





# Multilinguality

- Documents in a large number of languages

# Multilinguality

- Documents in a large number of languages
- Queries in a large number of languages

# Multilinguality

- Documents in a large number of languages
- Queries in a large number of languages
- First cut: Don't return English results for a Japanese query

# Multilinguality

- Documents in a large number of languages
- Queries in a large number of languages
- First cut: Don't return English results for a Japanese query
- However: Frequent mismatches query/document languages

# Multilinguality

- Documents in a large number of languages
- Queries in a large number of languages
- First cut: Don't return English results for a Japanese query
- However: Frequent mismatches query/document languages
- Many people can understand, but not query in a language

# Multilinguality

- Documents in a large number of languages
- Queries in a large number of languages
- First cut: Don't return English results for a Japanese query
- However: Frequent mismatches query/document languages
- Many people can understand, but not query in a language
- Translation is important.

# Multilinguality

- Documents in a large number of languages
- Queries in a large number of languages
- First cut: Don't return English results for a Japanese query
- However: Frequent mismatches query/document languages
- Many people can understand, but not query in a language
- Translation is important.
- Google example: "Beaujolais Nouveau -wine"

# Duplicate documents

- Significant duplication – 30%–40% duplicates in some studies



# Duplicate documents

- Significant duplication – 30%–40% duplicates in some studies
- Duplicates in the search results were common in the early days of the web.

# Duplicate documents

- Significant duplication – 30%–40% duplicates in some studies
- Duplicates in the search results were common in the early days of the web.
- Today's search engines eliminate duplicates very effectively.

# Duplicate documents

- Significant duplication – 30%–40% duplicates in some studies
- Duplicates in the search results were common in the early days of the web.
- Today's search engines eliminate duplicates very effectively.
- Key for high user satisfaction

# Trust

- For many collections, it is easy to assess the trustworthiness of a document.

# Trust

- For many collections, it is easy to assess the trustworthiness of a document.
  - A collection of Reuters newswire articles

# Trust

- For many collections, it is easy to assess the trustworthiness of a document.
  - A collection of Reuters newswire articles
  - A collection of TASS (Telegraph Agency of the Soviet Union) newswire articles from the 1980s

# Trust

- For many collections, it is easy to assess the trustworthiness of a document.
  - A collection of Reuters newswire articles
  - A collection of TASS (Telegraph Agency of the Soviet Union) newswire articles from the 1980s
  - Your Outlook email from the last three years

# Trust

- For many collections, it is easy to assess the trustworthiness of a document.
  - A collection of Reuters newswire articles
  - A collection of TASS (Telegraph Agency of the Soviet Union) newswire articles from the 1980s
  - Your Outlook email from the last three years
- Web documents are different: In many cases, we don't know how to evaluate the information.



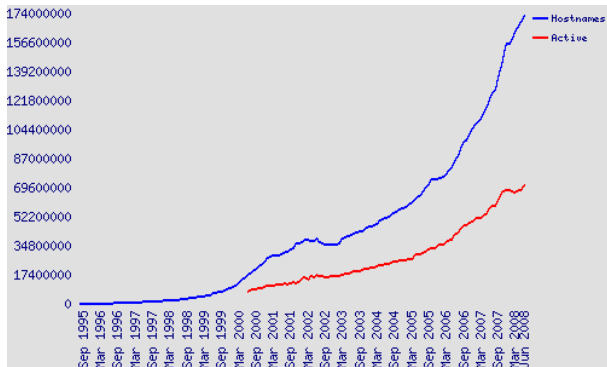
# Trust

- For many collections, it is easy to assess the trustworthiness of a document.
  - A collection of Reuters newswire articles
  - A collection of TASS (Telegraph Agency of the Soviet Union) newswire articles from the 1980s
  - Your Outlook email from the last three years
- Web documents are different: In many cases, we don't know how to evaluate the information.
- Hoaxes abound.

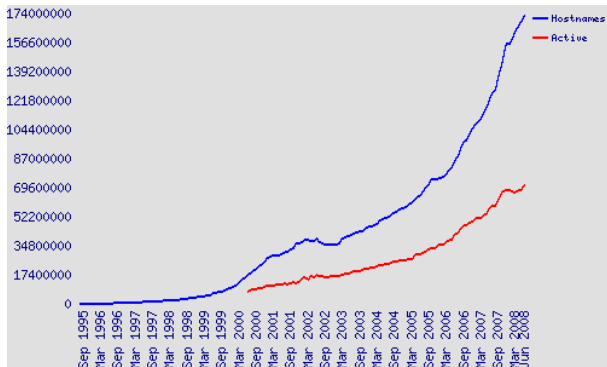
# Outline

- 1 Web IR
  - Links
  - Queries
  - Context
  - Users
  - Documents
  - Size
- 2 Ads & Spam
  - Ads
  - Spam

# Growth of the web

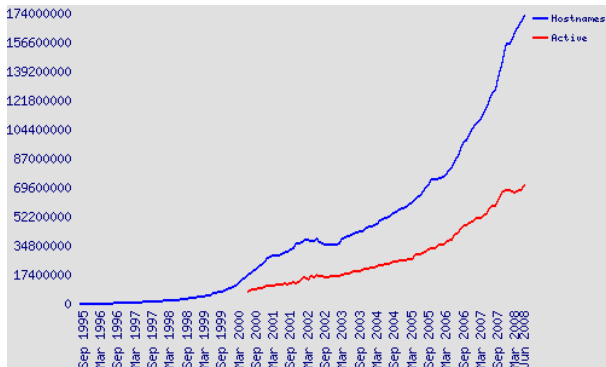


# Growth of the web



- The web keeps growing.

# Growth of the web



- The web keeps growing.
- But growth is no longer exponential?

# Size of the web: Issues

- What is size? Number of web servers? Number of pages?  
Terabytes of data available?

# Size of the web: Issues

- What is size? Number of web servers? Number of pages?  
Terabytes of data available?
- Some servers are seldom connected.

# Size of the web: Issues

- What is size? Number of web servers? Number of pages?  
Terabytes of data available?
- Some servers are seldom connected.
  - Example: Your laptop running a web server



# Size of the web: Issues

- What is size? Number of web servers? Number of pages?  
Terabytes of data available?
- Some servers are seldom connected.
  - Example: Your laptop running a web server
  - Is it part of the web?

# Size of the web: Issues

- What is size? Number of web servers? Number of pages?  
Terabytes of data available?
- Some servers are seldom connected.
  - Example: Your laptop running a web server
  - Is it part of the web?
- The “dynamic” web is infinite.

# Size of the web: Issues

- What is size? Number of web servers? Number of pages?  
Terabytes of data available?
- Some servers are seldom connected.
  - Example: Your laptop running a web server
  - Is it part of the web?
- The “dynamic” web is infinite.
  - Any sum of two numbers is its own dynamic page on Google.  
(Example: “2+4”)

## “Search engine index contains $N$ pages”: Issues

- Can I claim a page is in the index if I only index the first 4000 bytes?

## “Search engine index contains $N$ pages”: Issues

- Can I claim a page is in the index if I only index the first 4000 bytes?
- Can I claim a page is in the index if I only index anchor text pointing to the page?

## “Search engine index contains $N$ pages”: Issues

- Can I claim a page is in the index if I only index the first 4000 bytes?
- Can I claim a page is in the index if I only index anchor text pointing to the page?
  - There used to be (and still are?) billions of pages that are only indexed by anchor text.

How would you estimate the number of pages indexed by a web search engine?

## Simple method for determining a lower bound

- OR-query of frequent words in a number of languages



# Simple method for determining a lower bound

- OR-query of frequent words in a number of languages
- <http://ifnlp.org/ir/sizeoftheweb.html>

## Simple method for determining a lower bound

- OR-query of frequent words in a number of languages
- <http://ifnlp.org/ir/sizeoftheweb.html>
- According to this query: Size of web  $\geq 21,450,000,000$  on 2007.07.07 and  $\geq 25,350,000,000$  on 2008.07.03

## Simple method for determining a lower bound

- OR-query of frequent words in a number of languages
- <http://ifnlp.org/ir/sizeoftheweb.html>
- According to this query: Size of web  $\geq 21,450,000,000$  on 2007.07.07 and  $\geq 25,350,000,000$  on 2008.07.03
- But page counts of google search results are only rough estimates.

# Outline

- 1 Web IR
  - Links
  - Queries
  - Context
  - Users
  - Documents
  - Size
- 2 Ads & Spam
  - Ads
  - Spam

# Outline

- 1 Web IR
  - Links
  - Queries
  - Context
  - Users
  - Documents
  - Size
- 2 Ads & Spam
  - Ads
  - Spam

# First generation of search ads: Goto (1996)

www.goto.com/d/search/?sessionid=AAQ4214AAAAD0R50F0F3C6U0?type=home&time=18&keywords=Wilmington

**Wilmington real estate.**

Access 75% of all users now!  
Premium Listings reach 75% of all  
Internet users. [Go to](#) for Premium  
Listings today!

1. [Wilmington Real Estate - Buddy Blake](#)  
Wilmington's information and real estate guide. This is your on  
anything to do with Wilmington.  
[www.buddyblake.com](#) (Cost to advertiser: **\$0.25**)
2. [Coldwell Banker Sea Coast Realty](#)  
Wilmington's number one real estate company.  
[www.cbseacoast.com](#) (Cost to advertiser: [\\$0.22](#))
3. [Wilmington, NC Real Estate Becky Bullard](#)  
Everything you need to know about buying or selling a home c  
on my Web site!  
[www.iwvc.net](#) (Cost to advertiser: [\\$0.25](#))

# First generation of search ads: Goto (1996)

The screenshot shows a search results page from Goto.com. The URL in the browser's address bar is `www.goto.com/d/search/?sessionid=AAQ4214AAAAD050F0F3C6U0?type=home&time=18&keywords=Wilmington`. The page title is "Wilmington real estate." Below the title, there is a yellow box with the text: "Access 75% of all users now! Premium Listings reach 75% of all Internet users. [Go to](#) for Premium Listings today!". Below this, there is a list of three search results:

1. [Wilmington Real Estate - Buddy Blake](#)  
Wilmington's information and real estate guide. This is your on anything to do with Wilmington.  
[www.buddyblake.com](#) (Cost to advertiser: **\$0.28**)
2. [Coldwell Banker Sea Coast Realty](#)  
Wilmington's number one real estate company.  
[www.cbseacoast.com](#) (Cost to advertiser: **\$0.22**)
3. [Wilmington, NC Real Estate Becky Bullard](#)  
Everything you need to know about buying or selling a home c on my Web site!  
[www.iwvc.net](#) (Cost to advertiser: **\$0.25**)

- No separation of ads/docs. Just one result!

# First generation of search ads: Goto (1996)

www.goto.com/d/search/?sessionid=AAQ4214AAAADP50F0F3C6U0?type=home&tm=18&keywords=Wilmington

**Wilmington real estate.**

Access 75% of all users now!  
Premium Listings reach 75% of all  
Internet users. [Go to](#) for Premium  
Listings today!

1. [Wilmington Real Estate - Buddy Blake](#)  
Wilmington's information and real estate guide. This is your on  
anything to do with Wilmington.  
[www.buddyblake.com](#) (Cost to advertiser: **\$0.38**)
2. [Coldwell Banker Sea Coast Realty](#)  
Wilmington's number one real estate company.  
[www.cbseacoast.com](#) (Cost to advertiser: [\\$0.32](#))
3. [Wilmington, NC Real Estate Becky Bullard](#)  
Everything you need to know about buying or selling a home c  
on my Web site!  
[www.iwvc.net](#) (Cost to advertiser: [\\$0.35](#))

- No separation of ads/docs. Just one result!
- Buddy Blake bid the maximum (\$0.38) for this search.



# First generation of search ads: Goto (1996)

www.goto.com/d/search/?sessionid=AAQ4214AAAAD850F1F3C6A0?type=home&tr=1&keywords=Wilmington

**Wilmington real estate.**

Access 75% of all users now!  
Premium Listings reach 75% of all  
Internet users. [Go to](#) for Premium  
Listings today!

1. [Wilmington Real Estate - Buddy Blake](#)  
Wilmington's information and real estate guide. This is your on  
anything to do with Wilmington.  
[www.buddyblake.com](#) (Cost to advertiser: **\$0.38**)
2. [Coldwell Banker Sea Coast Realty](#)  
Wilmington's number one real estate company.  
[www.cbseacoast.com](#) (Cost to advertiser: **\$0.32**)
3. [Wilmington, NC Real Estate Becky Bullard](#)  
Everything you need to know about buying or selling a home c  
on my Web site!  
[www.iwvc.net](#) (Cost to advertiser: **\$0.35**)

- No separation of ads/docs. Just one result!
- Buddy Blake bid the maximum (\$0.38) for this search.
- He paid \$0.38 to Goto every time somebody clicked on the link.

# First generation of search ads: Goto (1996)

www.goto.com/d/search/?sessionid=AAQ4214AAAAD8150F1F3C6A0?type=home&time=18&keywords=Wilmington

**Wilmington real estate.**

Access 75% of all users now!  
Premium Listings reach 75% of all Internet users. [Go to](#) for Premium Listings today!

1. [Wilmington Real Estate - Buddy Blake](#)  
Wilmington's information and real estate guide. This is your on anything to do with Wilmington.  
[www.buddyblake.com](#) (Cost to advertiser: **\$0.38**)
2. [Coldwell Banker Sea Coast Realty](#)  
Wilmington's number one real estate company.  
[www.cbseacoast.com](#) (Cost to advertiser: [\\$0.32](#))
3. [Wilmington, NC Real Estate Becky Bullard](#)  
Everything you need to know about buying or selling a home c on my Web site!  
[www.iwvc.net](#) (Cost to advertiser: [\\$0.35](#))

- No separation of ads/docs. Just one result!
- Buddy Blake bid the maximum (\$0.38) for this search.
- He paid \$0.38 to Goto every time somebody clicked on the link.
- Upfront and honest. No relevance ranking, but Goto did not pretend there was any.

## Second generation of search ads: Google (2000/2001)

- Strict separation of search results and search ads

# Ranking of advertisers in search results

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 Commission\*, 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 Firstrate for Free!

[www.firstrate.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

# Ranking of advertisers in search results

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/](http://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 Commission", Reduced Commission, Account Fee Per Year (How to Avoid), Avg. ...

[www.smartmoney.com/brokers/index.cfm?story=2004-discount-table](http://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](http://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](http://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/](http://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](http://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 Firstrate for Free!  
[www.firstrate.com](http://www.firstrate.com)

## Discount Broker

Commission free trades for 30 days.  
No maintenance fees. Sign up now.  
[TDAMERITRADE.com](http://TDAMERITRADE.com)

## TradeKing - Online Broker

\$4.95 per Trade, Market or Limit  
SmartMoney Top **Discount Broker** 2007  
[www.TradeKing.com](http://www.TradeKing.com)

## Scottrade Brokerage

\$7 Trades, No Share Limit. In-Depth  
Research. Start Trading Online Now!  
[www.Scottrade.com](http://www.Scottrade.com)

## Stock trades \$1.00 to \$3


100 free trades, up to \$100 back  
for transfer costs, \$500 minimum  
[www.sogotrade.com](http://www.sogotrade.com)

## \$3.95 Online Stock Trades

Market/Limit Orders, No Share Limit  
and No Inactivity Fees  
[www.Marsco.com](http://www.Marsco.com)

## INGDIRECT | ShareBuilder

SogoTrade appears in ads.



# Ranking of advertisers in search results

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/](http://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 Commission\*, Reduced Commission, Account Fee Per Year (How to Avoid), Avg. ...

[www.smartmoney.com/brokers/index.cfm?story=2004-discount-table](http://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](http://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](http://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/](http://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](http://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 Firstrate for Free!  
[www.firstrate.com](http://www.firstrate.com)

## Discount Broker

Commission free trades for 30 days.  
No maintenance fees. Sign up now.  
[TDAMERITRADE.com](http://TDAMERITRADE.com)

## TradeKing - Online Broker

\$4.95 per Trade, Market or Limit  
SmartMoney Top **Discount Broker** 2007  
[www.TradeKing.com](http://www.TradeKing.com)

## Scottrade Brokerage

\$7 Trades, No Share Limit. In-Depth  
Research. Start Trading Online Now!  
[www.Scottrade.com](http://www.Scottrade.com)

## Stock trades \$1 to \$3

100 free trades, up to \$100 back  
for transfer costs, \$500 minimum  
[www.sogotrade.com](http://www.sogotrade.com)

## \$3.95 Online Stock Trades

Market/Limit Orders, No Share Limit  
and No Inactivity Fees  
[www.Marsco.com](http://www.Marsco.com)

## INGDIRECT | ShareBuilder

SogoTrade appears in search results.

SogoTrade appears in ads.

# Ranking of advertisers in search results

Web Images Maps News Shopping Gmail more

Sign in



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/](http://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 Commission\*, Reduced Commission, Account Fee Per Year (How to Avoid), Avg. ...

[www.smartmoney.com/brokers/index.cfm?story=2004-discount-table](http://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](http://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](http://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/](http://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](http://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 Firstrate for Free!  
[www.firstrate.com](http://www.firstrate.com)

## Discount Broker

Commission free trades for 30 days.  
No maintenance fees. Sign up now.  
[TDAMERITRADE.com](http://TDAMERITRADE.com)

## TradeKing - Online Broker

\$4.95 per Trade, Market or Limit  
SmartMoney Top **Discount Broker** 2007  
[www.TradeKing.com](http://www.TradeKing.com)

## Scotttrade Brokerage

\$7 Trades, No Share Limit. In-Depth  
Research. Start Trading Online Now!  
[www.Scotttrade.com](http://www.Scotttrade.com)

## Stock trades \$1.00 - \$3

100 free trades, up to \$100 back  
for transfer costs, \$500 minimum  
[www.sogotrade.com](http://www.sogotrade.com)

## \$3.95 Online Stock Trades

Market/Limit Orders, No Share Limit  
and No Inactivity Fees  
[www.Marsco.com](http://www.Marsco.com)

## INGDIRECT | ShareBuilder

SogoTrade appears in search results.

SogoTrade appears in ads.

Do search engines rank advertisers higher than non-advertisers?

# Ranking of advertisers in search results

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/](http://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 Commission", Reduced Commission, Account Fee Per Year (How to Avoid), Avg. ...

[www.smartmoney.com/brokers/index.cfm?story=2004-discount-table](http://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](http://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](http://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/](http://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](http://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 Firstrate for Free!  
[www.firstrate.com](http://www.firstrate.com)

## Discount Broker

Commission free trades for 30 days.  
No maintenance fees. Sign up now.  
[TDAMERITRADE.com](http://TDAMERITRADE.com)

## TradeKing - Online Broker

\$4.95 per Trade, Market or Limit  
SmartMoney Top **Discount Broker** 2007  
[www.TradeKing.com](http://www.TradeKing.com)

## Scotttrade Brokerage

\$7 Trades, No Share Limit, In-Depth  
Research. Start Trading Online Now!  
[www.Scotttrade.com](http://www.Scotttrade.com)

## Stock trades \$1 - \$3

100 free trades, up to \$100 back  
for transfer costs, \$500 minimum  
[www.sogotrade.com](http://www.sogotrade.com)

## \$3.95 Online Stock Trades

Market/Limit Orders, No Share Limit  
and No Inactivity Fees  
[www.Marsco.com](http://www.Marsco.com)

## INGDIRECT | ShareBuilder

SogoTrade appears in search results.

SogoTrade appears in ads.

Do search engines rank advertisers higher than non-advertisers?

All major search engines claim no.



# Do ads influence editorial content?

- Similar problem at newspapers / TV channels

# Do ads influence editorial content?

- Similar problem at newspapers / TV channels
- A newspaper is reluctant to publish harsh criticism of its major advertisers.

# Do ads influence editorial content?

- Similar problem at newspapers / TV channels
- A newspaper is reluctant to publish harsh criticism of its major advertisers.
- The line often gets blurred at newspapers / on TV.

# Do ads influence editorial content?

- Similar problem at newspapers / TV channels
- A newspaper is reluctant to publish harsh criticism of its major advertisers.
- The line often gets blurred at newspapers / on TV.
- No known case of this happening with search engines yet?

# How are ads placed?

- Advertisers bid for keywords.

# How are ads placed?

- Advertisers bid for keywords.
- Open system: Anybody can participate and bid on keywords.

# How are ads placed?

- Advertisers bid for keywords.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are only charged when somebody clicks on your ad.

# How are ads placed?

- Advertisers bid for keywords.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are only charged when somebody clicks on your ad.
- How does the advertiser determine its bid price and how does the search engine take this into account in ranking and billing?



# How are ads placed?

- Advertisers bid for keywords.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are only charged when somebody clicks on your ad.
- How does the advertiser determine its bid price and how does the search engine take this into account in ranking and billing?
  - Basis is a second price auction, but with twists

# How are ads placed?

- Advertisers bid for keywords.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are only charged when somebody clicks on your ad.
- How does the advertiser determine its bid price and how does the search engine take this into account in ranking and billing?
  - Basis is a second price auction, but with twists
  - Lots of interesting work on this

# How are ads placed?

- Advertisers bid for keywords.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are only charged when somebody clicks on your ad.
- How does the advertiser determine its bid price and how does the search engine take this into account in ranking and billing?
  - Basis is a second price auction, but with twists
  - Lots of interesting work on this
  - Squeeze an additional fraction of a cent from each ad means billions of additional revenue for the search engine.

# How are ads placed?

- Advertisers bid for keywords.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are only charged when somebody clicks on your ad.
- How does the advertiser determine its bid price and how does the search engine take this into account in ranking and billing?
  - Basis is a second price auction, but with twists
  - Lots of interesting work on this
  - Squeeze an additional fraction of a cent from each ad means billions of additional revenue for the search engine.
  - Many interesting jobs in this area

# Keywords with high bids

According to <http://www.cwire.org/highest-paying-search-terms/>

\$69.1	mesothelioma treatment options	\$35.86	pennsylvania medical malpractice attorney
\$65.85	personal injury lawyer michigan	\$35.86	medical malpractice ohio
\$62.59	student loans consolidation	\$35.71	automobile insurance quote
\$61.44	car accident attorney los angeles	\$35.4	loan consolidating
\$59.44	online car insurance quotes	\$35.34	commercial insurance quote
\$59.39	arizona dui lawyer	\$35.33	tax attorney
\$57.87	michigan car accident attorney	\$35.15	home equity loans
\$56.59	free auto insurance quote	\$34.81	instant auto insurance quotes
\$53.17	personal injury lawyers los angeles	\$34.8	home equity loan rates
\$52.31	free online auto insurance quote	\$34.79	home owners insurance quotes
\$50.4	accident attorney michigan	\$34.71	home equity line
\$50.35	michigan auto accident attorney	\$34.53	compensation solicitors
\$49.25	accident helpline	\$34.38	automobile insurance quotes
\$47.74	automobile accident lawyers	\$34.37	term insurance quotes
\$47.49	dui defense attorneys	\$34.26	instant car insurance quotes
\$46.44	asbestos cancer	\$34.02	auto insurance online quote
\$46.34	arizona dui	\$33.49	new york criminal attorney
\$45.8	business liability insurance quote	\$33.45	secured loan
\$43.86	loan consolidation	\$33.44	equity lines
\$42.98	student loan consolidation	\$33.41	criminal lawyer new york
\$40.7	dui defense lawyers	\$33.36	refinance mortgage
\$40.1	home equity line of credit	\$33.12	equity loan rates
\$39.81	life insurance quotes	\$33.07	manhattan mini storage
\$39.78	criminal lawyers new york	\$32.46	equity line
\$39.32	loan federal consolidation	\$32.45	home equity credit
\$39.23	refinancing	\$32.02	loan consolidate
\$38.72	equity line of credit	\$31.98	secured loan consolidation
\$37.96	lasik eye surgery new york city	\$31.93	laser hair removal new york city
\$37	2nd mortgage	\$31.51	home equity rates
\$35.9	free car insurance quote	\$31.37	free credit report com

# Google AdWords demo

# How are ads ranked?

- First cut: according to bid price

# How are ads ranked?

- First cut: according to bid price
  - Bad idea: open to abuse



# How are ads ranked?

- First cut: according to bid price
  - Bad idea: open to abuse
  - Example: query *accident* → ad *buy a new car*

# How are ads ranked?

- First cut: according to bid price
  - Bad idea: open to abuse
  - Example: query *accident* → ad *buy a new car*
- Instead: rank based on bid price **and** relevance

# How are ads ranked?

- First cut: according to bid price
  - Bad idea: open to abuse
  - Example: query *accident* → ad *buy a new car*
- Instead: rank based on bid price **and** relevance
- Key measure of ad relevance: clickthrough rate

# How are ads ranked?

- First cut: according to bid price
  - Bad idea: open to abuse
  - Example: query *accident* → ad *buy a new car*
- Instead: rank based on bid price **and** relevance
- Key measure of ad relevance: clickthrough rate
- Result: A non-relevant ad will be ranked low.

# How are ads ranked?

- First cut: according to bid price
  - Bad idea: open to abuse
  - Example: query *accident* → ad *buy a new car*
- Instead: rank based on bid price **and** relevance
- Key measure of ad relevance: clickthrough rate
- Result: A non-relevant ad will be ranked low.
  - Even if this decreases search engine revenue short-term

# How are ads ranked?

- First cut: according to bid price
  - Bad idea: open to abuse
  - Example: query *accident* → ad *buy a new car*
- Instead: rank based on bid price **and** relevance
- Key measure of ad relevance: clickthrough rate
- Result: A non-relevant ad will be ranked low.
  - Even if this decreases search engine revenue short-term
  - Hope: Overall acceptance of the system and overall revenue is maximized if users get useful information.

# How are ads ranked?

- First cut: according to bid price
  - Bad idea: open to abuse
  - Example: query *accident* → ad *buy a new car*
- Instead: rank based on bid price **and** relevance
- Key measure of ad relevance: clickthrough rate
- Result: A non-relevant ad will be ranked low.
  - Even if this decreases search engine revenue short-term
  - Hope: Overall acceptance of the system and overall revenue is maximized if users get useful information.
- Other ranking factors: location, time of day, quality and loading speed of landing page

# How are ads ranked?

- First cut: according to bid price
  - Bad idea: open to abuse
  - Example: query *accident* → ad *buy a new car*
- Instead: rank based on bid price **and** relevance
- Key measure of ad relevance: clickthrough rate
- Result: A non-relevant ad will be ranked low.
  - Even if this decreases search engine revenue short-term
  - Hope: Overall acceptance of the system and overall revenue is maximized if users get useful information.
- Other ranking factors: location, time of day, quality and loading speed of landing page
- The main factor of course is the query – and possibly earlier queries in the same session.



# Ranking of advertisers in search results

Web [Images](#) [Maps](#) [News](#) [Shopping](#) [Gmail](#) [more](#)

[Sign in](#)

Google

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 Commission\*, 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 Firstrate for Free!

[www.firstrate.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

# Search ads: A win-win-win?

- The [search engine](#) company gets revenue every time somebody clicks on an ad.

# Search ads: A win-win-win?

- The **search engine** company gets revenue every time somebody clicks on an ad.
- The **user** only clicks on an ad if they are interested in the ad.

# Search ads: A win-win-win?

- The **search engine** company gets revenue every time somebody clicks on an ad.
- The **user** only clicks on an ad if they are interested in the ad.
  - Search engines punish misleading and nonrelevant ads.

# Search ads: A win-win-win?

- The **search engine** company gets revenue every time somebody clicks on an ad.
- The **user** only clicks on an ad if they are interested in the ad.
  - Search engines punish misleading and nonrelevant ads.
  - As a result, users are often satisfied with what they find after clicking on an ad.

# Search ads: A win-win-win?

- The **search engine** company gets revenue every time somebody clicks on an ad.
- The **user** only clicks on an ad if they are interested in the ad.
  - Search engines punish misleading and nonrelevant ads.
  - As a result, users are often satisfied with what they find after clicking on an ad.
  - Being willing to pay for ads on a search engine is a quality signal (one of many) that users take into account.

# Search ads: A win-win-win?

- The **search engine** company gets revenue every time somebody clicks on an ad.
- The **user** only clicks on an ad if they are interested in the ad.
  - Search engines punish misleading and nonrelevant ads.
  - As a result, users are often satisfied with what they find after clicking on an ad.
  - Being willing to pay for ads on a search engine is a quality signal (one of many) that users take into account.
- The **advertiser** finds new customers in a cost-effective way.

# The appeal of search ads to advertisers

- Why is web search potentially more attractive for advertisers than TV spots, newspaper ads or radio spots?



# The appeal of search ads to advertisers

- Why is web search potentially more attractive for advertisers than TV spots, newspaper ads or radio spots?
- Someone who just searched for “Saturn Aura Sport Sedan” is infinitely more likely to buy one than a random person watching TV.

# The appeal of search ads to advertisers

- Why is web search potentially more attractive for advertisers than TV spots, newspaper ads or radio spots?
- Someone who just searched for “Saturn Aura Sport Sedan” is infinitely more likely to buy one than a random person watching TV.
- Most importantly, the advertiser only pays if the customer took an action indicating interest (i.e., clicking on the ad).

# But frequently it's not a win-win-win

- Example: keyword arbitrage

# But frequently it's not a win-win-win

- Example: keyword arbitrage
  - Buy a keyword at Google

# But frequently it's not a win-win-win

- Example: keyword arbitrage
  - Buy a keyword at Google
  - Then redirect traffic to a third party that is paying much more than you had to pay to Google

# But frequently it's not a win-win-win

- Example: keyword arbitrage
  - Buy a keyword at Google
  - Then redirect traffic to a third party that is paying much more than you had to pay to Google
  - This rarely makes sense for the user.

# But frequently it's not a win-win-win

- Example: keyword arbitrage
  - Buy a keyword at Google
  - Then redirect traffic to a third party that is paying much more than you had to pay to Google
  - This rarely makes sense for the user.
- Ad spammers keep inventing new tricks.

# But frequently it's not a win-win-win

- Example: keyword arbitrage
  - Buy a keyword at Google
  - Then redirect traffic to a third party that is paying much more than you had to pay to Google
  - This rarely makes sense for the user.
- Ad spammers keep inventing new tricks.
- The search engines need time to catch up with them.



# Who owns a search term?

- Example: geico

# Who owns a search term?

- Example: geico
- During part of 2005: The search term “geico” on Google was bought by competitors.

# Who owns a search term?

- Example: geico
- During part of 2005: The search term “geico” on Google was bought by competitors.
- Geico lost this case in the United States.

# Who owns a search term?

- Example: geico
- During part of 2005: The search term “geico” on Google was bought by competitors.
- Geico lost this case in the United States.
- Currently in the courts: Louis Vuitton case in Europe

# Who owns a search term?

- Example: geico
- During part of 2005: The search term “geico” on Google was bought by competitors.
- Geico lost this case in the United States.
- Currently in the courts: Louis Vuitton case in Europe
- See [http://google.com/tm\\_complaint.html](http://google.com/tm_complaint.html)

# Outline

- 1 Web IR
  - Links
  - Queries
  - Context
  - Users
  - Documents
  - Size
- 2 Ads & Spam
  - Ads
  - Spam

# The goal of spamming on the web

- You have a page that will generate lots of revenue for you if people visit it.

# The goal of spamming on the web

- You have a page that will generate lots of revenue for you if people visit it.
- Therefore, you would like to direct visitors to this page.



# The goal of spamming on the web

- You have a page that will generate lots of revenue for you if people visit it.
- Therefore, you would like to direct visitors to this page.
- One way of doing this: get your page ranked highly in search results.

# The goal of spamming on the web

- You have a page that will generate lots of revenue for you if people visit it.
- Therefore, you would like to direct visitors to this page.
- One way of doing this: get your page ranked highly in search results.
- How can I get my page ranked highly?

# Spam technique: Keyword stuffing / Hidden text

- Misleading meta-tags, excessive repetition

# Spam technique: Keyword stuffing / Hidden text

- Misleading meta-tags, excessive repetition
- Hidden text with colors, style sheet tricks etc.

# Spam technique: Keyword stuffing / Hidden text

- Misleading meta-tags, excessive repetition
- Hidden text with colors, style sheet tricks etc.
- Used to be very effective, most search engines now catch these



# Spam technique: Doorway and lander pages

- Doorway page: optimized for a single keyword, redirects to the real target page

# Spam technique: Doorway and lander pages

- Doorway page: optimized for a single keyword, redirects to the real target page
- Lander page: optimized for a single keyword or a misspelled domain name, designed to attract surfers who will then click on ads



# Lander page

Weitere Links: [Wild Yam Root](#) | [Mexican Appetizers](#) | [Yam](#) | [Gambar Skodeng Ulu Yam](#) | [Wild Eyes](#) | [The Yam Yams](#) | [Amica Cream](#) | [Chickweed Cream](#) | [Colloidal Silver Cream](#) | [Witch Hazel Cream](#) |

## COMPOSITA.COM

 Sprachauswahl: Deutsch ▾

### Sponsored Links

#### [Wild Russian Girls](#)

Plenty of Russian Girls interested in building a Happy Marriage.  
[uk.anastasia-international.com](http://uk.anastasia-international.com)

#### [Wild Yam 10%](#)

By HPLC , Supply 500Kg/mon from 100% natural herb  
[www.honsonbio.com](http://www.honsonbio.com)

#### [Suche dir eine Frau aus](#)

Sofort Kontakte zu Frauen Ohne Anmeldung, kostenlos starten!  
[www.SMS-Contacts.de/Sexy](http://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](http://rightboat.com/adverts/Yamaha.html)

#### [Wild Yam Root](#)

Harvested at height of potency. 20 Year, Family Run Herb Company.  
[www.BlessedHerbs.com](http://www.BlessedHerbs.com)

### WEITERE LINKS

- ▾ [Wild Yam Root](#)
- ▾ [Mexican Appetizers](#)
- ▾ [Yam](#)
- ▾ [Gambar Skodeng Ulu Yam](#)
- ▾ [Wild Eyes](#)
- ▾ [The Yam Yams](#)
- ▾ [Amica Cream](#)
- ▾ [Chickweed Cream](#)
- ▾ [Colloidal Silver Cream](#)
- ▾ [Witch Hazel Cream](#)

# Lander page

Weitere Links: [Wild Yam Root](#) | [Mexican Appetizers](#) | [Yam](#) | [Gambar Skodeng Ulu Yam](#) | [Wild Eyes](#) | [The Yam Yams](#) | [Amica Cream](#) | [Chickweed Cream](#) | [Colloidal Silver Cream](#) | [Witch Hazel Cream](#) |

## COMPOSITA.COM

 Sprachauswahl: Deutsch ▾

### Sponsored Links

#### [Wild Russian Girls](#)

Plenty of Russian Girls interested in building a Happy Marriage.  
[uk.anastasia-international.com](http://uk.anastasia-international.com)

#### [Wild Yam 10%](#)

By HPLC , Supply 500Kg/mon from 100% natural herb  
[www.honsonbio.com](http://www.honsonbio.com)

#### [Suche dir eine Frau aus](#)

Sofort Kontakte zu Frauen Ohne Anmeldung, kostenlos starten!  
[www.SMS-Contacts.de/Sexy](http://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](http://rightboat.com/adverts/Yamaha.html)

#### [Wild Yam Root](#)

Harvested at height of potency. 20 Year, Family Run Herb Company.  
[www.BlessedHerbs.com](http://www.BlessedHerbs.com)

### WEITERE LINKS

- ▾ [Wild Yam Root](#)
- ▾ [Mexican Appetizers](#)
- ▾ [Yam](#)
- ▾ [Gambar Skodeng Ulu Yam](#)
- ▾ [Wild Eyes](#)
- ▾ [The Yam Yams](#)
- ▾ [Amica Cream](#)
- ▾ [Chickweed Cream](#)
- ▾ [Colloidal Silver Cream](#)
- ▾ [Witch Hazel Cream](#)

- Number one hit on Google for the search “composita”

# Lander page

Weitere Links: [Wild Yam Root](#) | [Mexican Appetizers](#) | [Yam](#) | [Gambar Skodeng Ulu Yam](#) | [Wild Eyes](#) | [The Yam Yams](#) | [Amica Cream](#) | [Chickweed Cream](#) | [Colloidal Silver Cream](#) | [Witch Hazel Cream](#) |

## COMPOSITA.COM

 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](#)
- ▾ [Amica Cream](#)
- ▾ [Chickweed Cream](#)
- ▾ [Colloidal Silver Cream](#)
- ▾ [Witch Hazel Cream](#)

- Number one hit on Google for the search “composita”
- The only purpose of this page: get people to click on the ads and make money for the page owner

# Spam technique: Duplication

- Get good content from somewhere (steal it or produce it yourself)

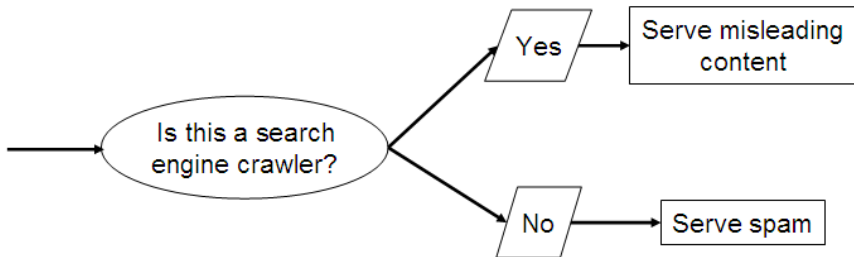
# Spam technique: Duplication

- Get good content from somewhere (steal it or produce it yourself)
- Publish a large number of slight variations of it

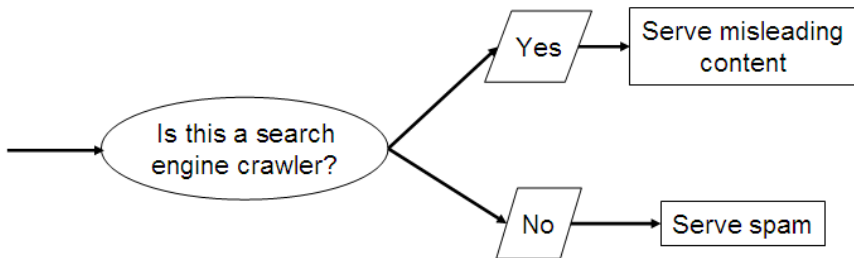
# Spam technique: Duplication

- Get good content from somewhere (steal it or produce it yourself)
- Publish a large number of slight variations of it
- For example, publish the answer to a tax question with the spelling variations of “tax deferred” on the previous slide

# Spam technique: Cloaking



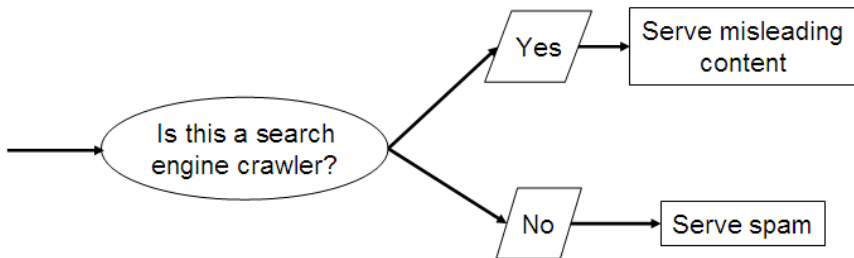
# Spam technique: Cloaking



- Serve fake content to search engine spider

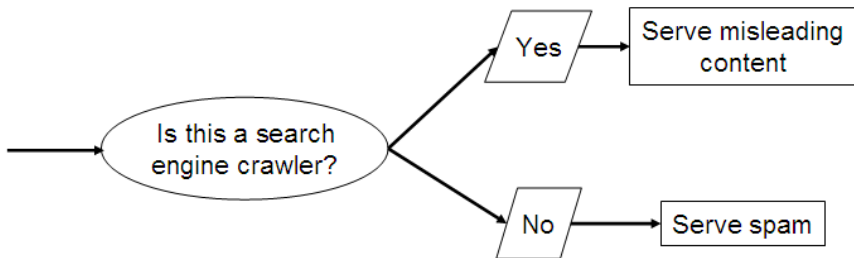


# Spam technique: Cloaking



- Serve fake content to search engine spider
- So do we just penalize this always?

# Spam technique: Cloaking



- Serve fake content to search engine spider
- So do we just penalize this always?
- No: legitimate uses (e.g., different content to US vs. European users)

# Spam technique: Link spam

- Create lots of links pointing to the page you want to promote

# Spam technique: Link spam

- Create lots of links pointing to the page you want to promote
- Put these links on pages with high (or at least non-zero) PageRank

# Spam technique: Link spam

- Create lots of links pointing to the page you want to promote
- Put these links on pages with high (or at least non-zero) PageRank
  - Newly registered domains (domain flooding)

# Spam technique: Link spam

- Create lots of links pointing to the page you want to promote
- Put these links on pages with high (or at least non-zero) PageRank
  - Newly registered domains (domain flooding)
  - A set of pages that all point to each other to boost each other's PageRank (mutual admiration society)

# Spam technique: Link spam

- Create lots of links pointing to the page you want to promote
- Put these links on pages with high (or at least non-zero) PageRank
  - Newly registered domains (domain flooding)
  - A set of pages that all point to each other to boost each other's PageRank (mutual admiration society)
  - Pay somebody to put your link on their highly ranked page (“schuetze horoskop” example)

# Spam technique: Link spam

- Create lots of links pointing to the page you want to promote
- Put these links on pages with high (or at least non-zero) PageRank
  - Newly registered domains (domain flooding)
  - A set of pages that all point to each other to boost each other's PageRank (mutual admiration society)
  - Pay somebody to put your link on their highly ranked page (“schuetze horoskop” example)
  - Leave comments that include the link on blogs



# SEO: Search engine optimization

- Promoting a page in the search rankings is not necessarily spam.

# SEO: Search engine optimization

- Promoting a page in the search rankings is not necessarily spam.
- It can also be a legitimate business – which is called SEO.

# SEO: Search engine optimization

- Promoting a page in the search rankings is not necessarily spam.
- It can also be a legitimate business – which is called SEO.
- You can hire an SEO firm to get your page highly ranked.

# SEO: Search engine optimization

- Promoting a page in the search rankings is not necessarily spam.
- It can also be a legitimate business – which is called SEO.
- You can hire an SEO firm to get your page highly ranked.
- There are many legitimate reasons for doing this.

# SEO: Search engine optimization

- Promoting a page in the search rankings is not necessarily spam.
- It can also be a legitimate business – which is called SEO.
- You can hire an SEO firm to get your page highly ranked.
- There are many legitimate reasons for doing this.
  - For example, Google bombs like *Who is a failure?*

# SEO: Search engine optimization

- Promoting a page in the search rankings is not necessarily spam.
- It can also be a legitimate business – which is called SEO.
- You can hire an SEO firm to get your page highly ranked.
- There are many legitimate reasons for doing this.
  - For example, Google bombs like *Who is a failure?*
- And there are many legitimate ways of achieving this:

# SEO: Search engine optimization

- Promoting a page in the search rankings is not necessarily spam.
- It can also be a legitimate business – which is called SEO.
- You can hire an SEO firm to get your page highly ranked.
- There are many legitimate reasons for doing this.
  - For example, Google bombs like *Who is a failure?*
- And there are many legitimate ways of achieving this:
  - Restructure your content in a way that makes it easy to index

# SEO: Search engine optimization

- Promoting a page in the search rankings is not necessarily spam.
- It can also be a legitimate business – which is called SEO.
- You can hire an SEO firm to get your page highly ranked.
- There are many legitimate reasons for doing this.
  - For example, Google bombs like *Who is a failure?*
- And there are many legitimate ways of achieving this:
  - Restructure your content in a way that makes it easy to index
  - Talk with influential bloggers and have them link to your site



# SEO: Search engine optimization

- Promoting a page in the search rankings is not necessarily spam.
- It can also be a legitimate business – which is called SEO.
- You can hire an SEO firm to get your page highly ranked.
- There are many legitimate reasons for doing this.
  - For example, Google bombs like *Who is a failure?*
- And there are many legitimate ways of achieving this:
  - Restructure your content in a way that makes it easy to index
  - Talk with influential bloggers and have them link to your site
  - Add more interesting and original content

# The war against spam

- Quality indicators

# The war against spam

- Quality indicators
  - Links, statistically analyzed (PageRank etc)

# The war against spam

- Quality indicators
  - Links, statistically analyzed (PageRank etc)
  - Usage (users visiting a page)

# The war against spam

- Quality indicators
  - Links, statistically analyzed (PageRank etc)
  - Usage (users visiting a page)
  - No adult content (e.g., no pictures with flesh-tone)

# The war against spam

- Quality indicators
  - Links, statistically analyzed (PageRank etc)
  - Usage (users visiting a page)
  - No adult content (e.g., no pictures with flesh-tone)
  - Distribution and structure of text (e.g., no keyword stuffing)

# The war against spam

- Quality indicators
  - Links, statistically analyzed (PageRank etc)
  - Usage (users visiting a page)
  - No adult content (e.g., no pictures with flesh-tone)
  - Distribution and structure of text (e.g., no keyword stuffing)
- Combine all of these indicators and use machine learning

# The war against spam

- Quality indicators
  - Links, statistically analyzed (PageRank etc)
  - Usage (users visiting a page)
  - No adult content (e.g., no pictures with flesh-tone)
  - Distribution and structure of text (e.g., no keyword stuffing)
- Combine all of these indicators and use machine learning
- Editorial intervention



# The war against spam

- Quality indicators
  - Links, statistically analyzed (PageRank etc)
  - Usage (users visiting a page)
  - No adult content (e.g., no pictures with flesh-tone)
  - Distribution and structure of text (e.g., no keyword stuffing)
- Combine all of these indicators and use machine learning
- Editorial intervention
  - Blacklists

# The war against spam

- Quality indicators
  - Links, statistically analyzed (PageRank etc)
  - Usage (users visiting a page)
  - No adult content (e.g., no pictures with flesh-tone)
  - Distribution and structure of text (e.g., no keyword stuffing)
- Combine all of these indicators and use machine learning
- Editorial intervention
  - Blacklists
  - Top queries audited

# The war against spam

- Quality indicators
  - Links, statistically analyzed (PageRank etc)
  - Usage (users visiting a page)
  - No adult content (e.g., no pictures with flesh-tone)
  - Distribution and structure of text (e.g., no keyword stuffing)
- Combine all of these indicators and use machine learning
- Editorial intervention
  - Blacklists
  - Top queries audited
  - Complaints addressed

# The war against spam

- Quality indicators
  - Links, statistically analyzed (PageRank etc)
  - Usage (users visiting a page)
  - No adult content (e.g., no pictures with flesh-tone)
  - Distribution and structure of text (e.g., no keyword stuffing)
- Combine all of these indicators and use machine learning
- Editorial intervention
  - Blacklists
  - Top queries audited
  - Complaints addressed
  - Suspect patterns detected

# Webmaster guidelines

- Major search engines have guidelines for webmasters.

# Webmaster guidelines

- Major search engines have guidelines for webmasters.
- These guidelines tell you what is legitimate SEO and what is spamming.

# Webmaster guidelines

- Major search engines have guidelines for webmasters.
- These guidelines tell you what is legitimate SEO and what is spamming.
- Ignore these guidelines at your own risk

# Webmaster guidelines

- Major search engines have guidelines for webmasters.
- These guidelines tell you what is legitimate SEO and what is spamming.
- Ignore these guidelines at your own risk
- Once a search engine identifies you as a spammer, all pages on your site may get low ranks (or disappear from the index entirely).



# Webmaster guidelines

- Major search engines have guidelines for webmasters.
- These guidelines tell you what is legitimate SEO and what is spamming.
- Ignore these guidelines at your own risk
- Once a search engine identifies you as a spammer, all pages on your site may get low ranks (or disappear from the index entirely).
- There is often a fine line between spam and legitimate SEO.

# Webmaster guidelines

- Major search engines have guidelines for webmasters.
- These guidelines tell you what is legitimate SEO and what is spamming.
- Ignore these guidelines at your own risk
- Once a search engine identifies you as a spammer, all pages on your site may get low ranks (or disappear from the index entirely).
- There is often a fine line between spam and legitimate SEO.
- Scientific study of fighting spam on the web: *adversarial information retrieval*

# Resources

- Chapter 19 of IIR

# Resources

- Chapter 19 of IIR
- Resources at <http://ifnlp.org/ir>

# Resources

- Chapter 19 of IIR
- Resources at <http://ifnlp.org/ir>
- Size of the web queries

# Resources

- Chapter 19 of IIR
- Resources at <http://ifnlp.org/ir>
- Size of the web queries
- Trademark issues (Geico and Vuitton cases)

# Resources

- Chapter 19 of IIR
- Resources at <http://ifnlp.org/ir>
- Size of the web queries
- Trademark issues (Geico and Vuitton cases)
- How ads are priced

# Resources

- Chapter 19 of IIR
- Resources at <http://ifnlp.org/ir>
- Size of the web queries
- Trademark issues (Geico and Vuitton cases)
- How ads are priced
- How search engines fight webspam



# Resources

- Chapter 19 of IIR
- Resources at <http://ifnlp.org/ir>
- Size of the web queries
- Trademark issues (Geico and Vuitton cases)
- How ads are priced
- How search engines fight webspam
- Adversarial IR site at Lehigh