

Information Retrieval on the Web

PRI 21/22 · Information Processing and Retrieval
M.EIC · Master in Informatics Engineering and Computation

Sérgio Nunes
Dept. Informatics Engineering
FEUP · U.Porto

Based on Chapters 19, 20 and 21 from Introduction to Information Retrieval, Manning, C. et al. (2008)
Based on Chapters 11 and 12 from Modern Information Retrieval (2nd Edition), Baeza-Yates, R. et al. (2010)

Outline

- Overview of Information Retrieval on the Web
- Web Crawling
- Link Analysis
 - PageRank
 - HITS

Information Retrieval on the Web

The World Wide Web

- The web is unprecedented in many ways:
 - Unprecedented in scale (size and change);
 - Unprecedented in lack of central coordination;
 - Unprecedented in the diversity of users' backgrounds and needs.
- Two types of challenges:
 - Data: distribution, size, volatility, quality, unstructured, duplicates.
 - Interaction: user needs; relevance; diversity of users.

Information Access on the Web

- Early project of making information discoverable on the web fell into two categories.
- **Full-text index search engines**, e.g. Altavista, Excite, Infoseek
 - Users use keyword-based search supported by inverted indexes and ranking mechanisms;
- **Directories / Taxonomies**, e.g. Yahoo!, Open Directory Project
 - Users navigate through a hierarchical tree of category labels.
 - Problems: mostly manual categorization (high cost and not scalable); mismatch between editors' and users' idea of how to organize a given node.

Web Search Challenges

- Decentralization of content publication (the web biggest innovation) means that there is no central point that keeps track of what exists, what was changed, or deleted.
- What is the size of the web?
- Content is created massively and in diverse forms
 - Diversity of languages and dialects;
 - Diversity of formats, i.e. structure, color, size, ...;
 - Quality, i.e. truths, lies, contradictions, ...;
 - Intent, i.e. legitimate, spam, search engine manipulation;

Web Characteristics

- The Web can be modeled as a graph.
- Web pages point to, and are pointed by, other pages.
- Links to other pages (out-links) usually include an "anchor text".
- The number of in-links to a page is called the in-degree.
- Studies of web characteristics and dynamics is an area of research.
- Early research suggests that the directed graph connecting web pages has a bowtie shape.

Web Graph

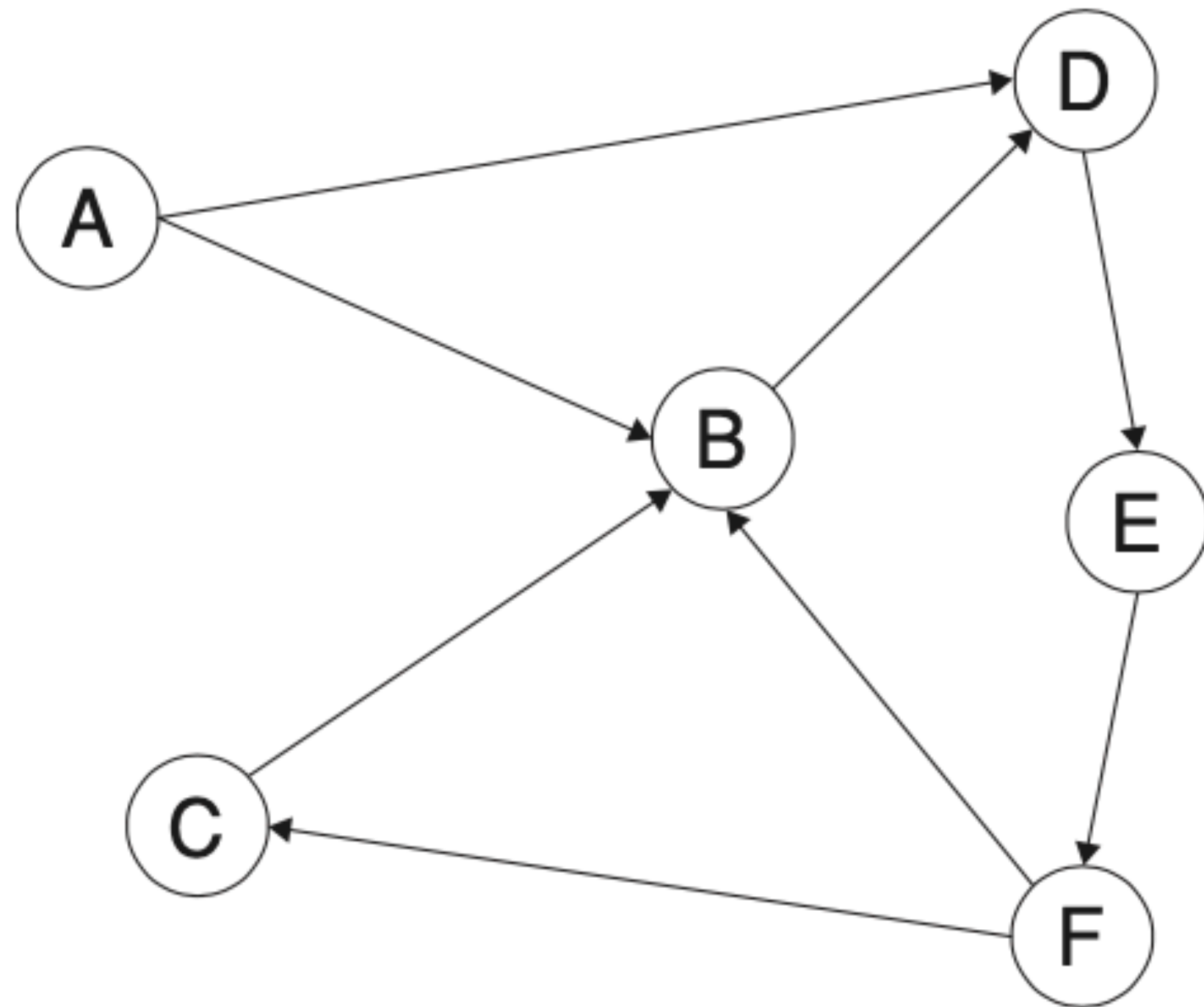
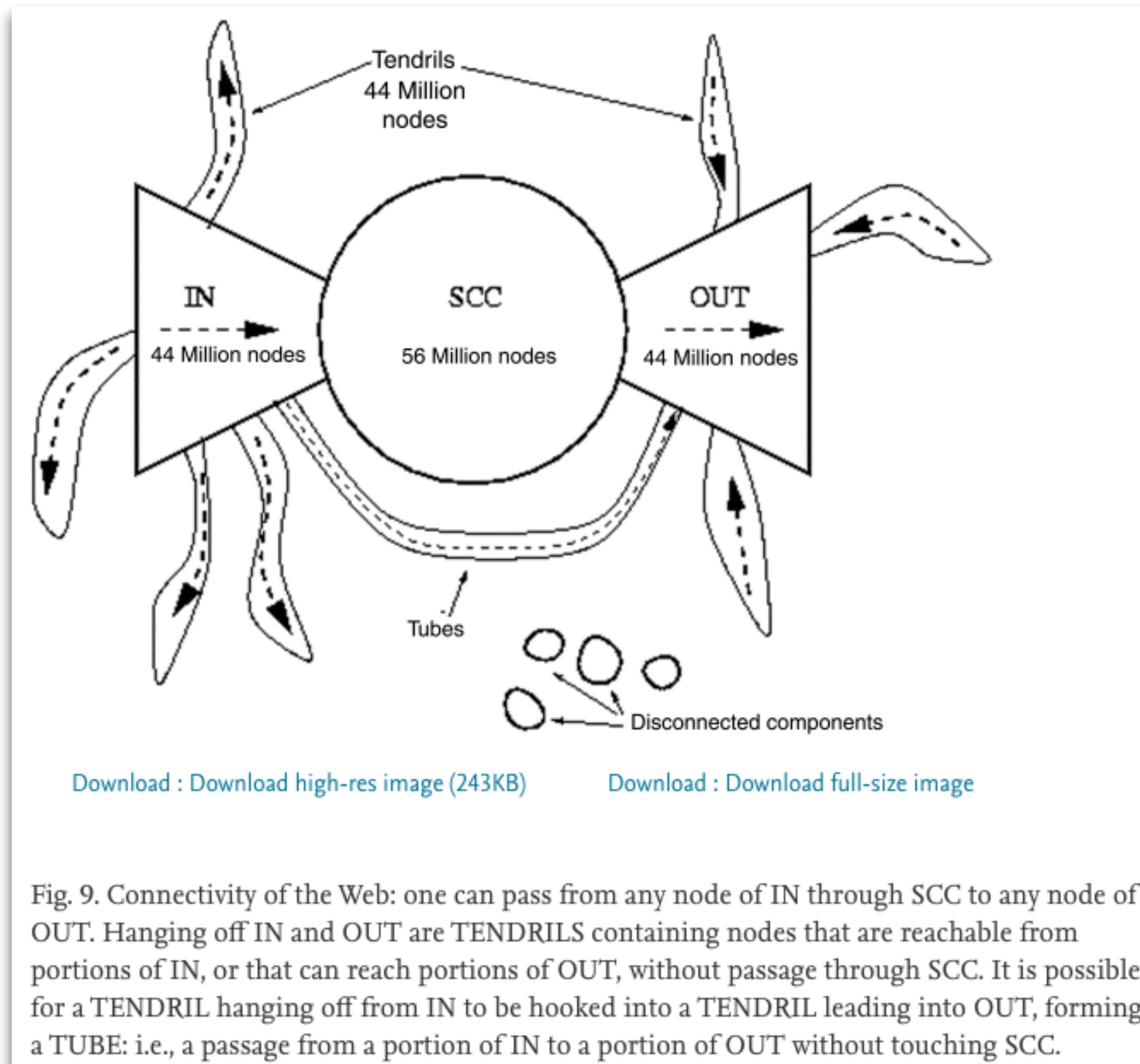


Figure 19.3 A sample small web graph. In this example we have six pages labeled A through F. Page B has in-degree 3 and out-degree 1. This example graph is not strongly connected: There is no path from any of pages B through F to page A.

Bowtie Shape of the Web



Web Spam

- There is a (high) commercial value associated with appearing on the top ranked results for a given search.
- For instance, a search engine whose scoring depends on the frequency of keywords, would be easy to manipulate by including numerous repetitions of selected keywords.
- This is called web spam, i.e. the manipulation of content on the web with the purpose of manipulating search engine rankings. Examples include: cloaking, link farms, link spam, click spam, etc.
- Topics in the sub-area of adversarial information retrieval.

Cloaking

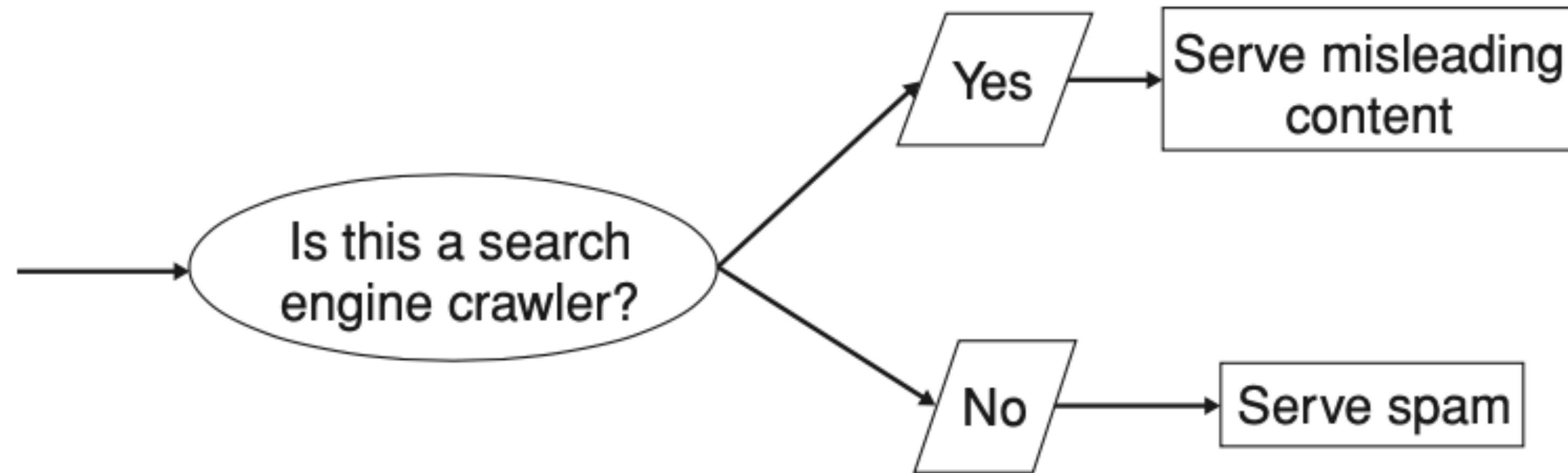


Figure 19.5 Cloaking as used by spammers.

"Mixed Motives in Search"

→ Reflexion from Brin and Page in 1998.

8 Appendix A: Advertising and Mixed Motives

Currently, the predominant business model for commercial search engines is advertising. The goals of the advertising business model do not always correspond to providing quality search to users. For example, in our prototype search engine one of the top results for cellular phone is "[The Effect of Cellular Phone Use Upon Driver Attention](#)", a study which explains in great detail the distractions and risk associated with conversing on a cell phone while driving. This search result came up first because of its high importance as judged by the PageRank algorithm, an approximation of citation importance on the web [[Page, 98](#)]. It is clear that a search engine which was taking money for showing cellular phone ads would have difficulty justifying the page that our system returned to its paying advertisers. For this type of reason and historical experience with other media [[Bagdikian 83](#)], we expect that advertising funded search engines will be inherently biased towards the advertisers and away from the needs of the consumers.

Since it is very difficult even for experts to evaluate search engines, search engine bias is particularly insidious. A good example was OpenText, which was reported to be selling companies the right to be listed at the top of the search results for particular queries [[Marchiori 97](#)]. This type of bias is much more insidious than advertising, because it is not clear who "deserves" to be there, and who is willing to pay money to be listed. This business model resulted in an uproar, and OpenText has ceased to be a viable search engine. But less blatant bias are likely to be tolerated by the market. For example, a search engine could add a small factor to search results from "friendly" companies, and subtract a factor from results from competitors. This type of bias is very difficult to detect but could still have a significant effect on the market. Furthermore, advertising income often provides an incentive to provide poor quality search results. For example, we noticed a major search engine would not return a large airline's homepage when the airline's name was given as a query. It so happened that the airline had placed an expensive ad, linked to the query that was its name. A better search engine would not have required this ad, and possibly resulted in the loss of the revenue from the airline to the search engine. In general, it could be argued from the consumer point of view that the better the search engine is, the fewer advertisements will be needed for the consumer to find what they want. This of course erodes the advertising supported business model of the existing search engines. However, there will always be money from advertisers who want a customer to switch products, or have something that is genuinely new. But we believe the issue of advertising causes enough mixed incentives that it is crucial to have a competitive search engine that is transparent and in the academic realm.

Search Engines Economic Model

- In the early years, banner advertisements were used to support search engines operation, i.e. static banners. These banners are typically priced on a cost per mil (CPM) basis, i.e. cost per 1,000 impressions.
- Alternatively, commercial contracts can be set on a number of clicks basis — cost per click (CPC). The advertiser only pays when the searcher clicks on a search result.
- Current economic models are based on a bidding paradigm, where advertisers bid on specific keywords and search engines show the ads based on the value of the bid.
- These results are known as sponsored search results, while the search engine "pure" ranking are known as algorithmic results.
- Targeting ads to keywords is a basic targeting strategy. More sophisticated (and problematic) approaches can explore contextual or behavioral user data, e.g. location, navigation patterns, previous searches, etc.

www.google.com/search?q=lego&biw=1111&bih=850

Google

lego

X

Entrar

Tudo

Imagens

Compras

Videos

Noticias

Mais

Ferramentas

Cerca de 1 090 000 000 resultados (0,77 segundos)

Anúncio · https://www.lego.com/

Conjuntos e exclusivos LEGO® - Compra conjuntos LEGO

Compra conjuntos e brinquedos **LEGO** e descobre o presente perfeito para as crianças. A maior seleção **LEGO** online. Explora agora em **LEGO.com**! Ganhe recompensas VIP. Qualidade garantida. Retorno sem complicações. Brands: Creator, Star Wars, Ideas.

Fender® Stratocaster™

Vive o sonho de estrela de rock com a guitarra Fender Strat! Compra já

Compra este Super Mario™

Revive memórias enquanto constróis o Bloco de interrogação Mario 64™

Compra o LEGO® Queer Eye

Constrói o Loft dos Fab Five da série galardoada Queer Eye

Anúncio · https://www.worten.pt/

Worten - Lego Construções - Aproveita Grandes Descontos

Entregas em Todo o País! Aproveita Já as Promoções em **Lego** Construções em Worten.pt. -50% Jogos e Brinquedos · Lego Friends · Lego Disney · Lego Minecraft · Lego Star Wars

Anúncio · https://www.toysrus.pt/

Jogos d Sets. Ampla coleção da LEGO | ToysRUs - ToysRUs.pt

Envio grátis sem compra mínima. Em todo o site | Toys"R"Us. Por tempo limitado. Se...

https://www.lego.com › pt-pt

Home | Loja LEGO® Oficial PT

16/12/2020 — Explore the world of **LEGO**® through games, videos, products and more! Shop awesome **LEGO**® building toys and brick sets and find the perfect ...

Anúncios · Comprar

LEGO Still Life with...
24,99 €
LEGO.com
★★★★★ (3)
De Google

LEGO Hogwarts...
429,99 €
LEGO.com
★★★★★ (1k+)
De Google

LEGO Camp Nou – FC...
329,99 €
LEGO.com
De Google

LEGO Infinity...
74,99 €
LEGO.com
★★★★★ (337)
De Google

PROMOÇÃO
LEGO CLASSIC
14,99 €
De Google

LEGO Ideas
De Google

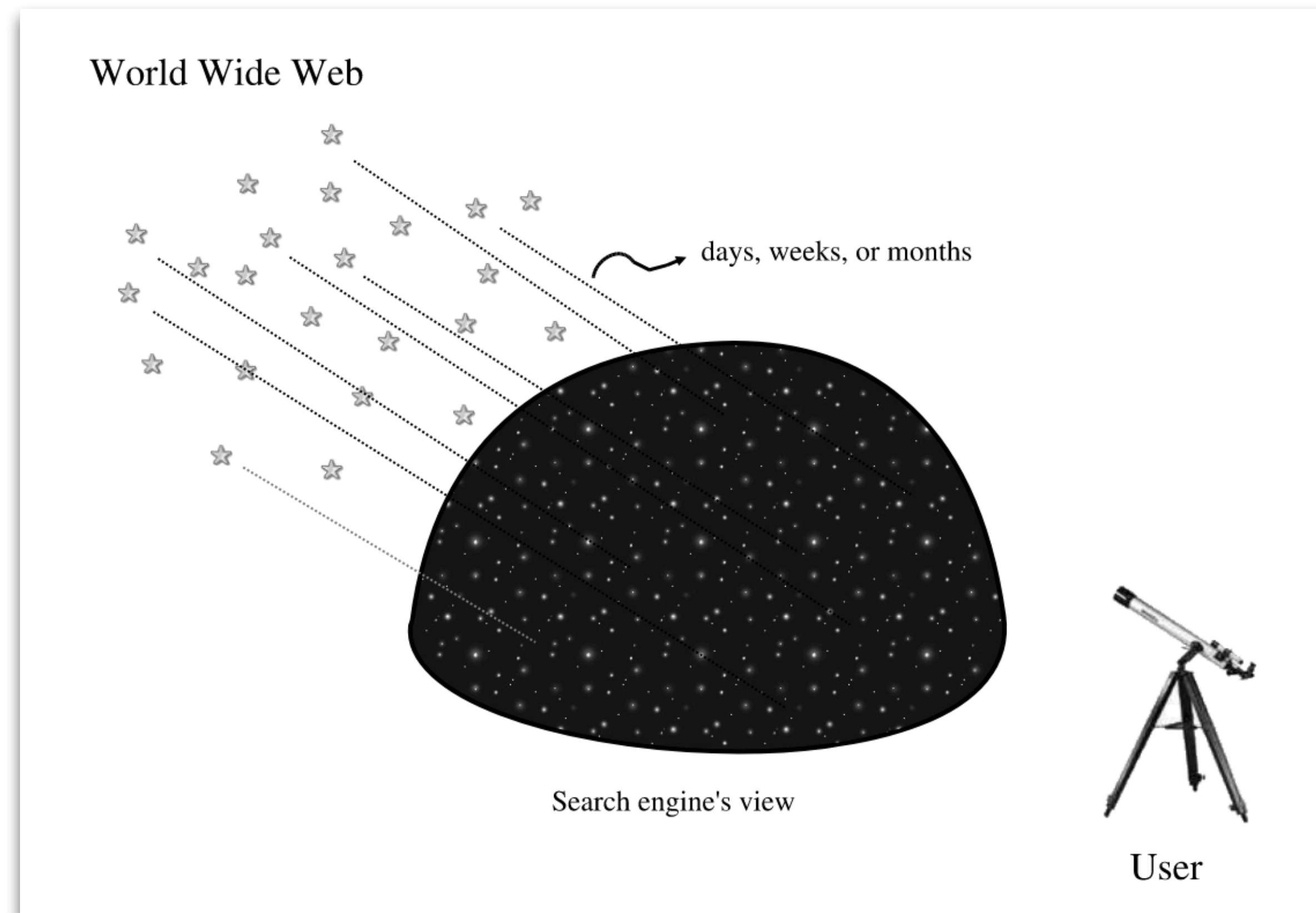
User Characteristics

- Web search users are not trained on how to write queries or the search operators offered by search engines.
- Studies in search user behavior show that users use between two or three keywords, search operators are rarely used, precision in the top results is highly valued, and lightweight result pages are preferred.
- Early research has shown that users' information needs can be grouped in three types:
 - Informational queries, i.e. seek general information about a topic. There is typically not a single source of relevant information. Users typically gather information from multiple web pages.
 - Navigational queries, i.e. seek the website or home page of a single entity. Users' expectations is to find a specific resource.
 - Transactional queries, i.e. preludes the user performing a transaction on the web, such as purchasing a product, downloading a file, or doing a reservation.

Signals for Web Ranking

- Hundreds of signals are used by search engines to estimate quality.
- Signals can be groups in different dimensions:
 - Query-independent signals (static).
 - Query-dependent signals (dynamic).
- Document-based signals (content or structural), e.g. HTML.
- Collection-based signals, e.g. Links.
- User-based signals, e.g. Clicks.

Web Crawling



Crawling the Web, in a certain way, resembles watching the sky in a clear night: the star positions that we see reflects the state of the stars at different times

Search Engine Architecture

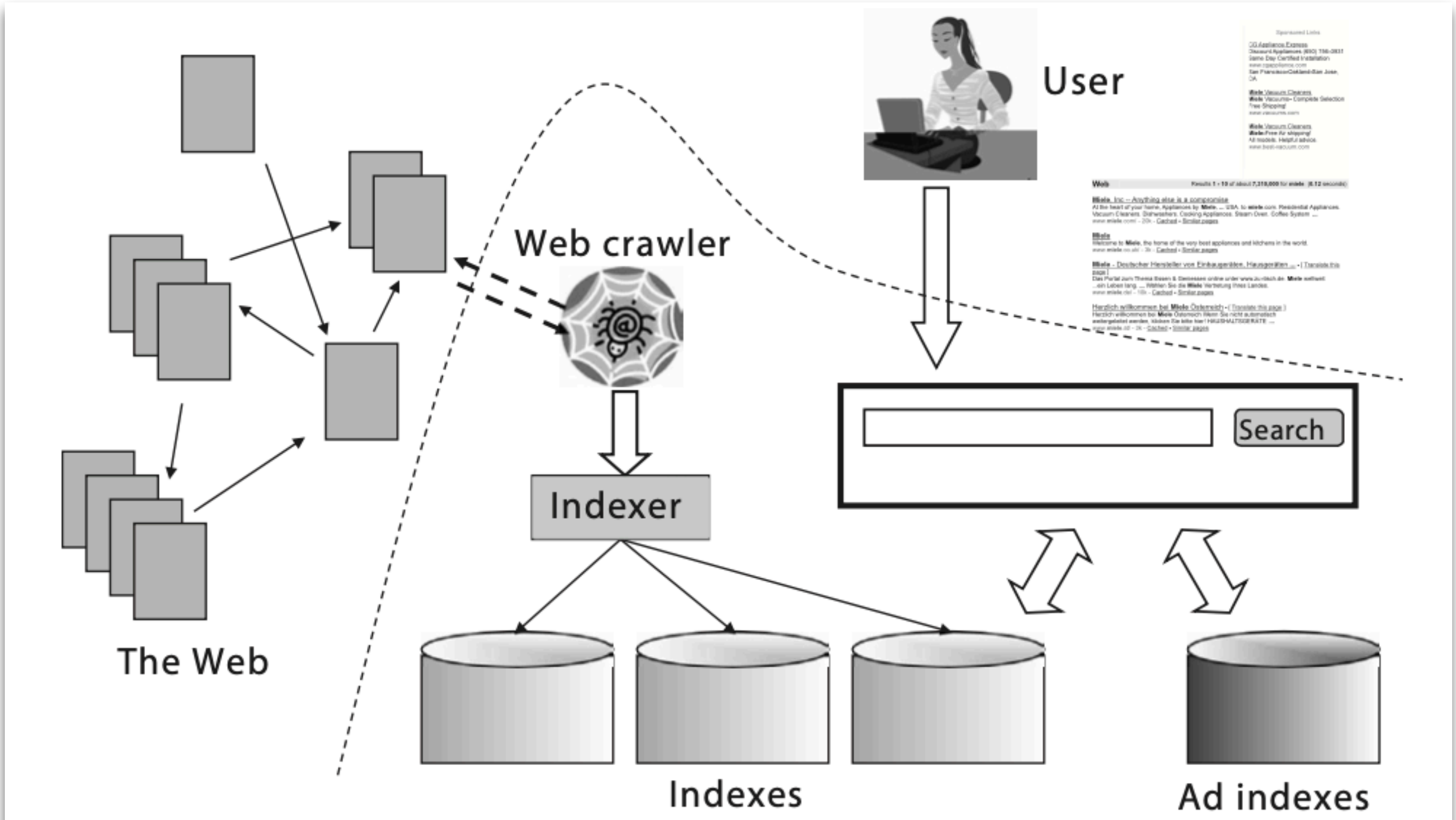


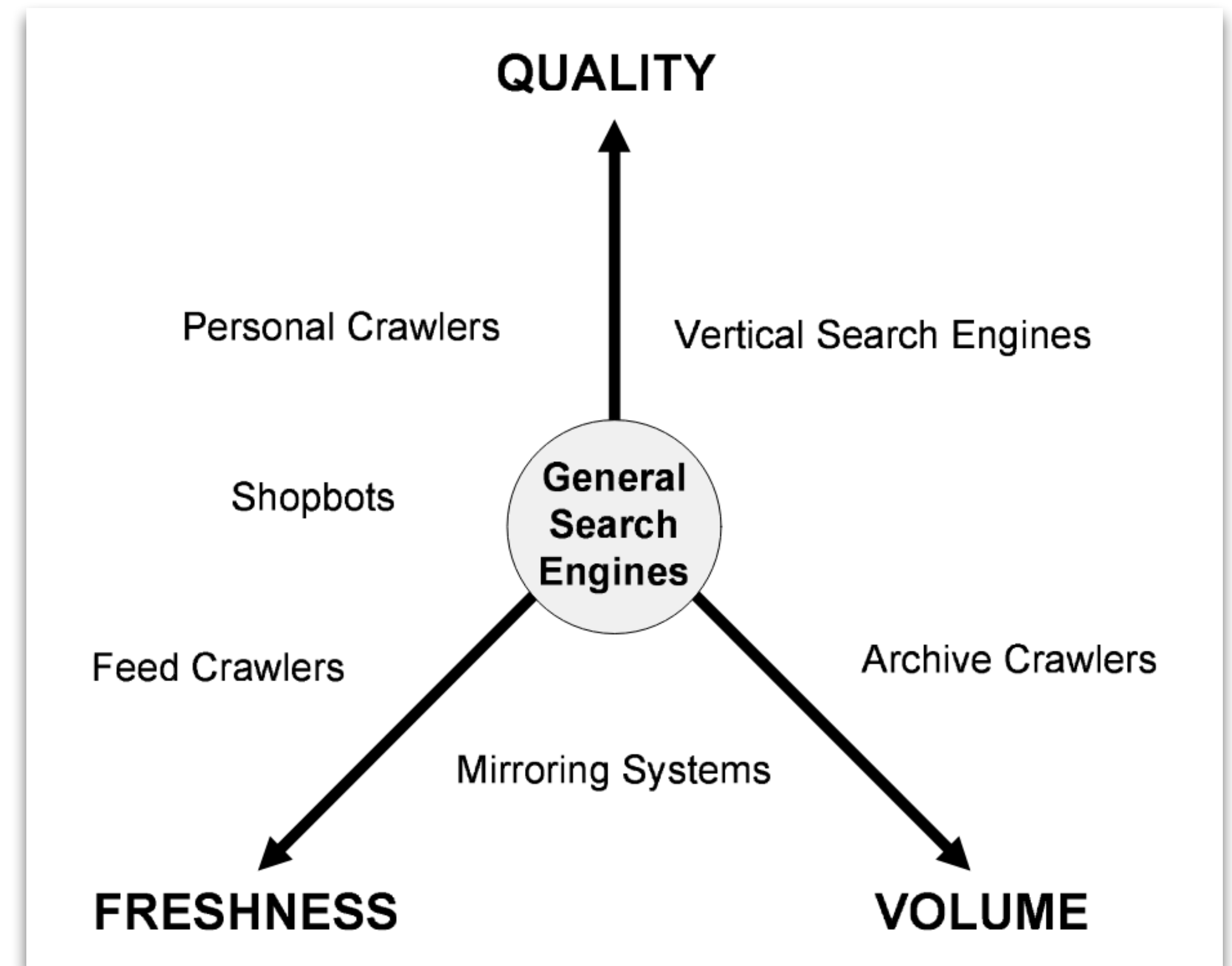
Figure 19.7 The various components of a web search engine.

Web Crawling

- Web crawling is the process by which pages are gathered from the web.
- The goal is to discover as quickly and as efficiently as possible quality web pages.
- Features a crawler **must** provide: robustness in face of problems or traps, politeness to web hosts.
- Features a crawler **should** provide: execute in a distributed fashion, scalable, efficient use of resources, bias towards good quality pages, freshness depending on page change rate, and extensible to cope with innovations on the web.
- Industry standards: robots.txt (robotstxt.org); sitemap protocol (sitemaps.org).

Taxonomy of Crawling

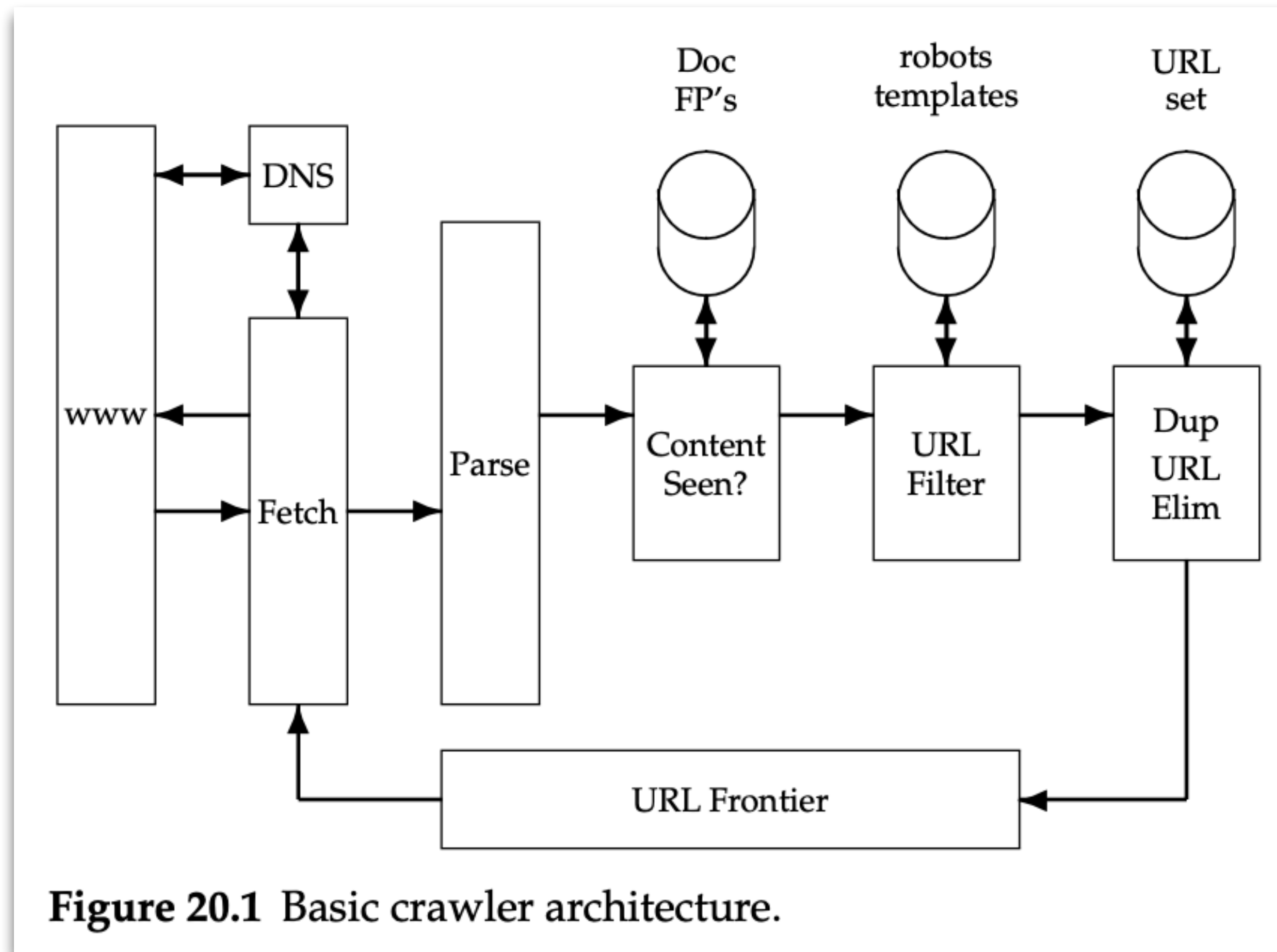
- The crawlers assign different importance to issues such as freshness, quality, and volume
- The crawlers can be classified according to these three axes



Taxonomy of Web Pages

		Dynamic	
		Reachable by following links	Reachable only by filling forms
Private	Static	Not Indexable	
	Public	Indexable by most search engines	Hidden/Deep Web

Web Crawler Architecture



Near-Duplicate Detection

- A large percentage of content on the web are near-duplicates, i.e. differing only in small parts. Standard duplicate detection methods do not work, e.g. fingerprinting.
- Crawlers need to decide if new pages are duplicates of existing content and if pages being revisited have changed since last visit, e.g. estimate change rate.
- Common solution: obtain shingles of k size from web pages; compare shingles from two pages to determine if they are near-duplicates. The more shingles in common, the more similar are the pages.
- Example of $k=4$ shingles for the text: [these are red and blue roses]
 - [these are red and], [are red and blue], [red and blue roses]

Open Source Crawlers

- Heritrix, Internet Archive (Java), <https://github.com/internetarchive/heritrix3>
- Nutch, Apache (Java), <https://nutch.apache.org/>
- Scrapy, (Python), <https://github.com/scrapy/scrapy>

Link Analysis

Link-based Signals

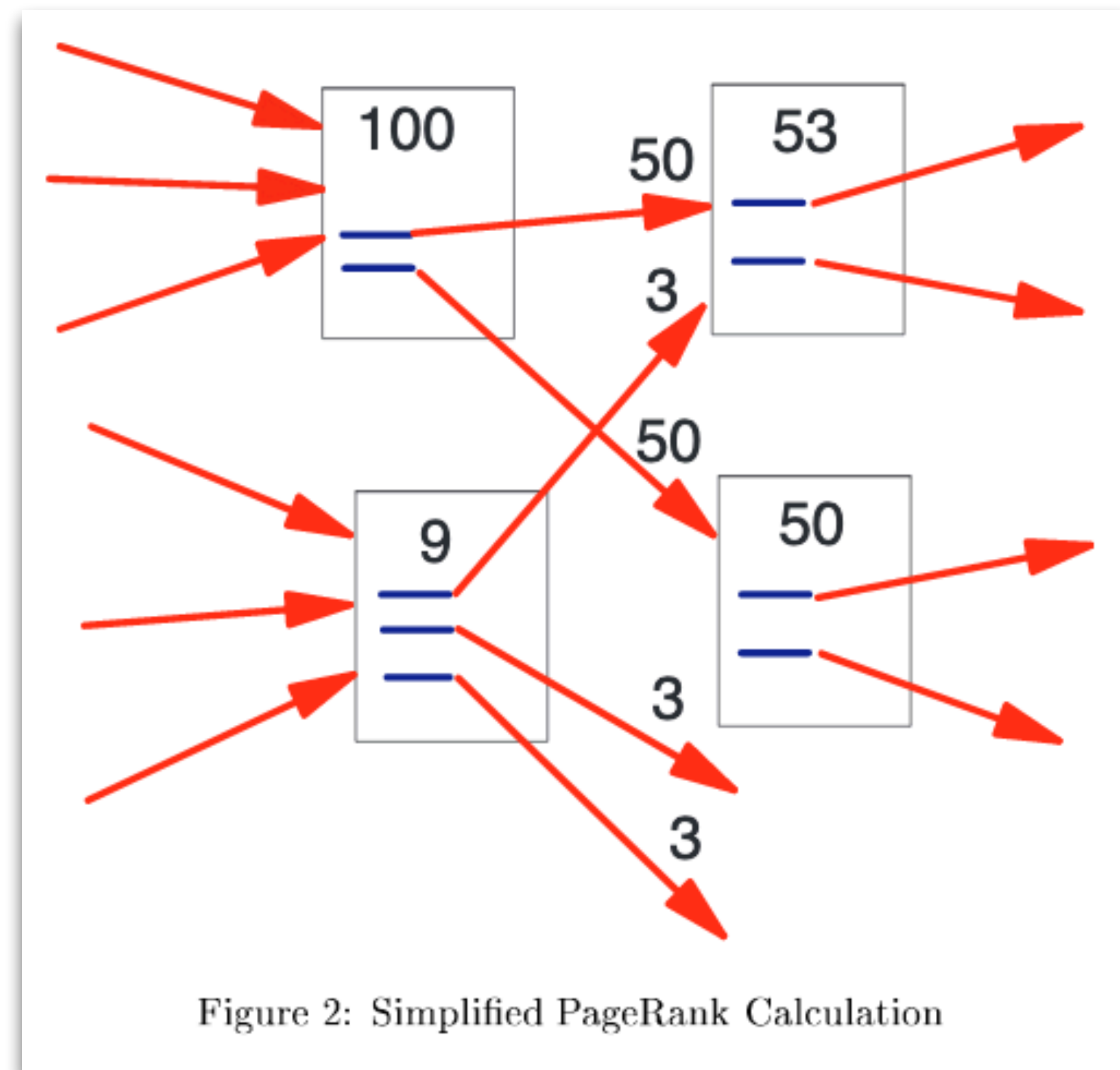
- Links are one of the distinctive features of a collection of web documents.
- Base assumption: the number of hyperlinks pointing to a page provides a measure of its popularity and quality.
- Link-based ranking algorithms build on the assumption that an hyperlink from page A to page B represents an endorsement of page B, by the creator of page A.
- Two classic algorithms: PageRank and HITS.
 - PageRank, Larry Page and Sergey Brin (1996)
 - HITS, Jon Kleinberg (1997)

PageRank

- The PageRank of a node is a value between 0 and 1.
- It is a query-independent value computed offline that only depends on the structure of the web graph.
- The algorithm models a random surfer who begins at a web page and, at each step, randomly chooses between the out-links to move to the next page. If the random surfer executes this "forever", he will visit some pages more frequently than others. The PageRank value of a page represents this probability.

$$PR(a) = \frac{q}{T} + (1 - q) \sum_{i=1}^n \frac{PR(p_i)}{L(p_i)}$$

PageRank Illustration

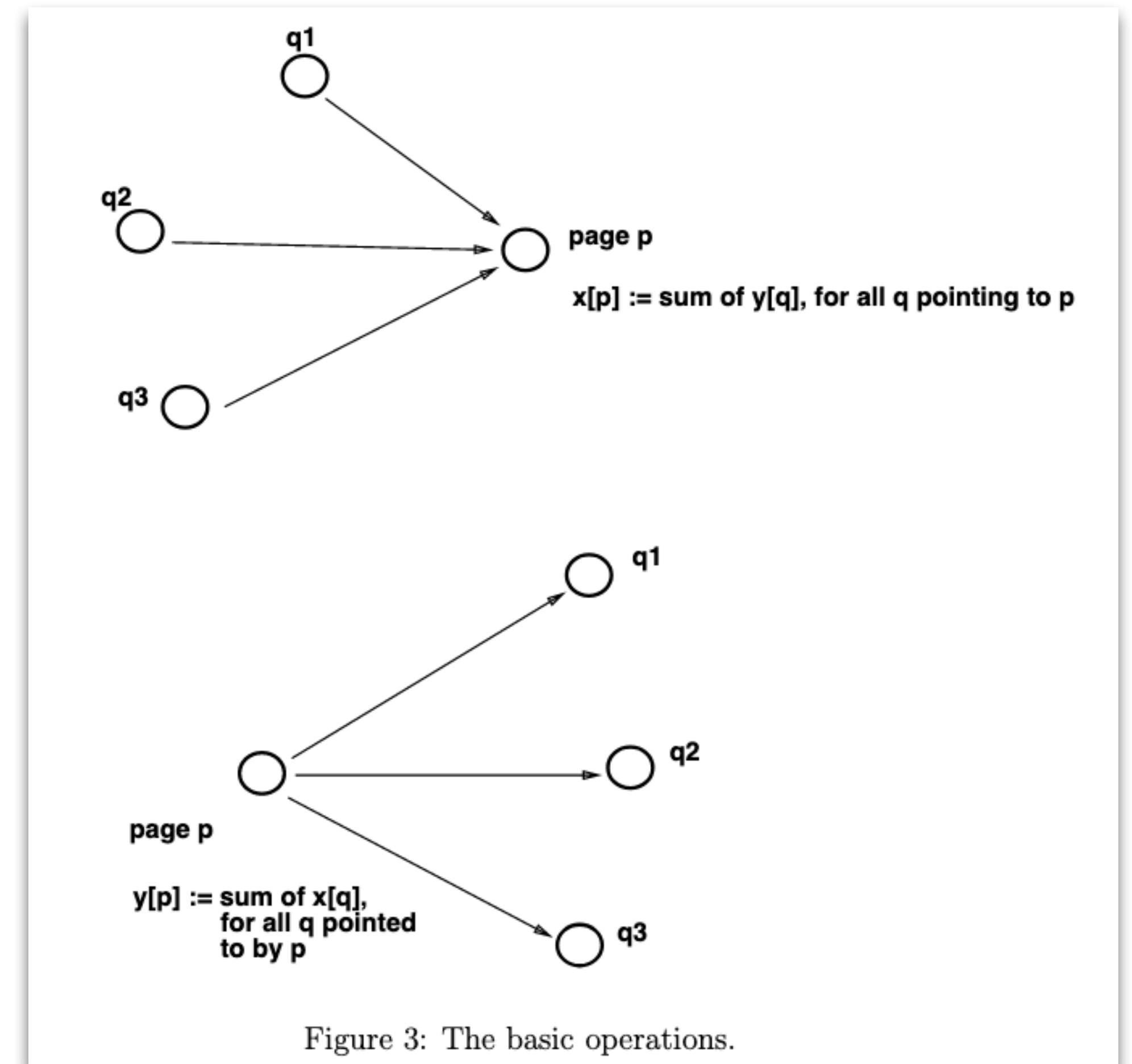
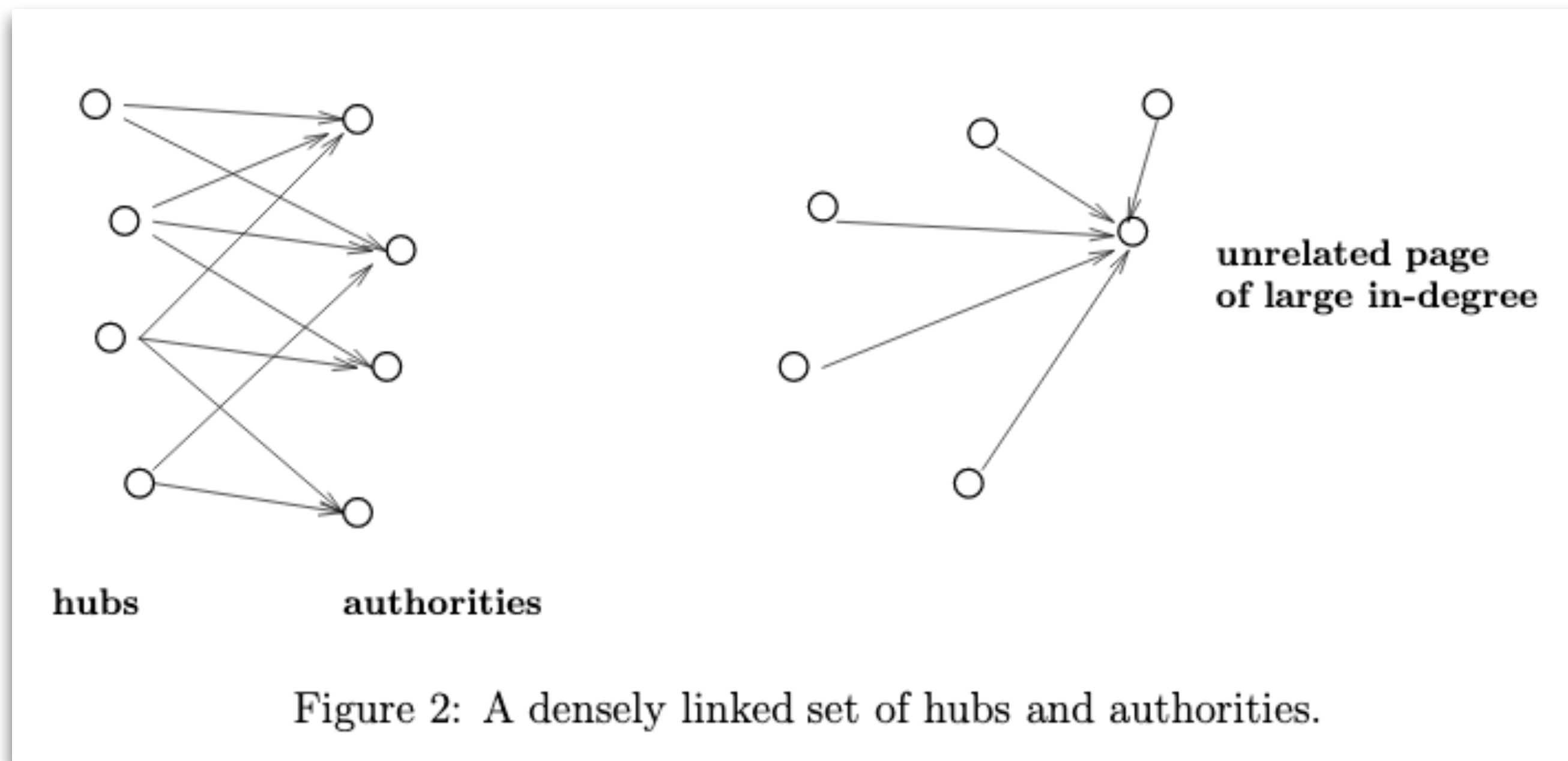


HITS — Hyperlink Induced Topic Search

- Query-dependent algorithm.
- Starts with the answer set (e.g. pages containing the keywords).
- Computes two scores for each page: authority score and hub score.
 - Pages with many links pointing to them are called authorities.
 - Pages with many outgoing links are called hubs.

$$H(p) = \sum_{u \in S \mid p \rightarrow u} A(u) , \quad A(p) = \sum_{v \in S \mid v \rightarrow p} H(v)$$

HITS Illustration



Anchor Text as a Signal

- The text used in HTML anchors, i.e. links, is called anchor text.
- `Faculty of Engineering of the University of Porto`
- Represents a description from "others" about a given web page.
- The collection of all anchor texts can be explored with standard IR techniques, and incorporated as an additional features in an inverted index.
- Important feature for image search.
- See "Google bombing", https://en.wikipedia.org/wiki/Google_bombing

References

- Sergey Brin and Lawrence Page, The anatomy of a large-scale hypertextual Web search engine. Comput. Netw. ISDN Syst. 30, 1-7 (April 1998), <http://infolab.stanford.edu/~backrub/google.html>
- Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval: The Concepts and Technology behind Search
 - [Chapter 11: Web Retrieval](#) + [Slides for Chapter 11: Web Retrieval](#)
 - [Slides for Chapter 12: Web Crawling](#)
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval
 - [Chapter 19: Web search basics](#)
 - [Chapter 20: Web crawling and indexes](#)
 - [Chapter 21: Link analysis](#)