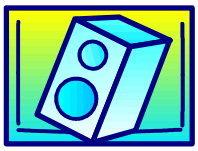


Institut für Rundfunkökonomie
an der Universität zu Köln

**Denis Türker**

# The Optimal Design of a Search Engine from an Agency Theory Perspective

**Arbeitspapiere**
**des Instituts für Rundfunkökonomie**
**an der Universität zu Köln**

**Heft Nr. 191**

**Köln, im August 2004**

**Arbeitspapiere des Instituts für Rundfunkökonomie**

**Working Papers of the Institute for Broadcasting Economics**

ISSN der Arbeitspapiere: 0945-8999
*ISSN of the Working Papers: 0945-8999*

ISBN des vorliegenden Arbeitspapiers 189: 3-934156-85-1
*ISBN of the Working Paper at hand 189:* 3-934156-85-1

Schutzgebühr 14,-- €
*Price 14,-- €*

Die Arbeitspapiere können im Internet eingesehen und abgerufen werden unter der Adresse
http://www.rundfunk-institut.uni-koeln.de

*The Working Papers can be read and downloaded from the Internet URL*
*http://www.rundfunk-institut.uni-koeln.de*

*Mitteilungen und Bestellungen richten Sie bitte per E-Mail an:*
*rundfunk-institut@uni-koeln.de*
*oder an die unten genannte Postanschrift*

*Messages and Orders to the Institute can be sent via Email to:*
rundfunk-institut@uni-koeln.de
*or to the mailing address mentioned below.*

**Institut für Rundfunkökonomie**
**an der Universität zu Köln**

Hohenstaufenring 57a
D-50674 Köln
Telefon: (0221) 23 35 36
Telefax: (0221) 24 11 34

**Denis Türker**

# The Optimal Design of a Search Engine from an Agency Theory Perspective*



* Slighly modified version of a thesis that was accepted by the economic faculty of the University of Cologne in spring term 2004.

## List of Figures

## Abbreviations

| | |
|---|---|
| AI | Artificial Intelligence |
| API | Application Program Interface |
| DMOZ | Open Directory Project |
| DSL | Digital Subscriber Line |
| HITS | Hypertext Induced Topic Search |
| IA | Intermediary Agent |
| IDF | Inverse Document Frequency |
| IE | Information Extraction |
| IPO | Initial Public Offering |
| IR | Information Retrieval |
| NLP | Natural Language Processing |
| OCR | Optical Character Recognition |
| QA | Question Answering |
| TF | Term Frequency |
| URL | Uniform Resource Locator |
| WWW | World Wide Web |

# 1. Introduction

In today's world, information is arguably the most important resource. Search engines have emerged as powerful gatekeepers, indispensable tools in our efforts to manage and access the wealth of material available online. The 'perfect' search engine would intuitively comprehend a user's desire for information and retrieve the most relevant documents in response to even poorly formulated queries.

Efforts to design the definitive search engine can be compared to a winemaker's pursuit of the perfect wine. Only the best grapes, picked at just the right moment, can be crafted into an excellent wine. Decisions made in the vineyard, winery, and cellar all affect the final product. Even if a winegrower is satisfied with his work, a sophisticated vinophile may feel differently. Moreover, other factors, such as the meal with which a wine is served, may influence preferences. By the same logic, an individual searching the Web for both a local restaurant and technical scientific papers might get better results employing different search engines, each designed to meet those specific needs.

What else do an outstanding wine and a superior search engine have in common? A winemaker's decisions are comparable to a search service provider's choices regarding indexed documents (grapes harvested), ranking techniques (fermentation), and search results (bottling). Both products require a supply of the best input factors and production techniques, while product choice is largely dependent upon user needs. An effective search engine is built on high-quality, updated, and unbiased content.

On the other hand, there are major differences between search engines and exceptional wines. Most search engine users are not willing to pay. Consequently, many providers accept advertising to gain revenues. In contrast to fine wines, information (the traded good of search engines) is certainly not scarce. Digital goods can be sold and kept at the same time; originals and copies cannot be distinguished. The currency of the Internet is the limited attention of users. Search engines gain revenues by directing this attention to ad clients.

This paper outlines four common search engine design patterns that determine a service provider's business model, field of activity, and relationships with stakeholders. These strategies are based on technical aspects and trends in search engine architecture (chapters 2 and 3), characteristics of the information market and media industries (chapter 4), and tenets of agency theory that elucidate the relationships between stakeholders (chapter 5).

# 2. Web Information Retrieval

The focus of this paper is to define an optimal search engine from an Agency Theory perspective. Basic technical knowledge is required to address these issues. This chapter provides an overview of well-known information retrieval (IR) models, introduces different forms of web search, and examines the relationship between web pages and search engines.

## 2.1. Overviews of IR-Models

Traditional IR systems make use of index terms to retrieve documents. The query is keyword-based and simple, but has a significant disadvantage, in that the intended meaning of terms can be lost in the process.[1] A ranking algorithm determines which documents are relevant, with those regarded to be more relevant listed first. The following figure shows a classification of IR Models suggested by Baeza-Yates and Ribeiro-Neto.[2]

### 2.1.1. Retrieval: Ad Hoc and Filtering

*Ad hoc* retrieval is the operational mode for conventional IR systems, wherein most documents of the collection remain relatively static while new queries are submitted to the system.

The operational mode of *filtering* is employed for collections that change while relatively static queries are submitted. An example is the case of a stock market. A profile, describing the preferences of the user, is used to filter the incoming documents. This approach is also used for the selection of news articles that might be of interest to the user. The main task of filtering is not the ranking of documents, but the creation of profiles that represent the preferences of the user.[3]

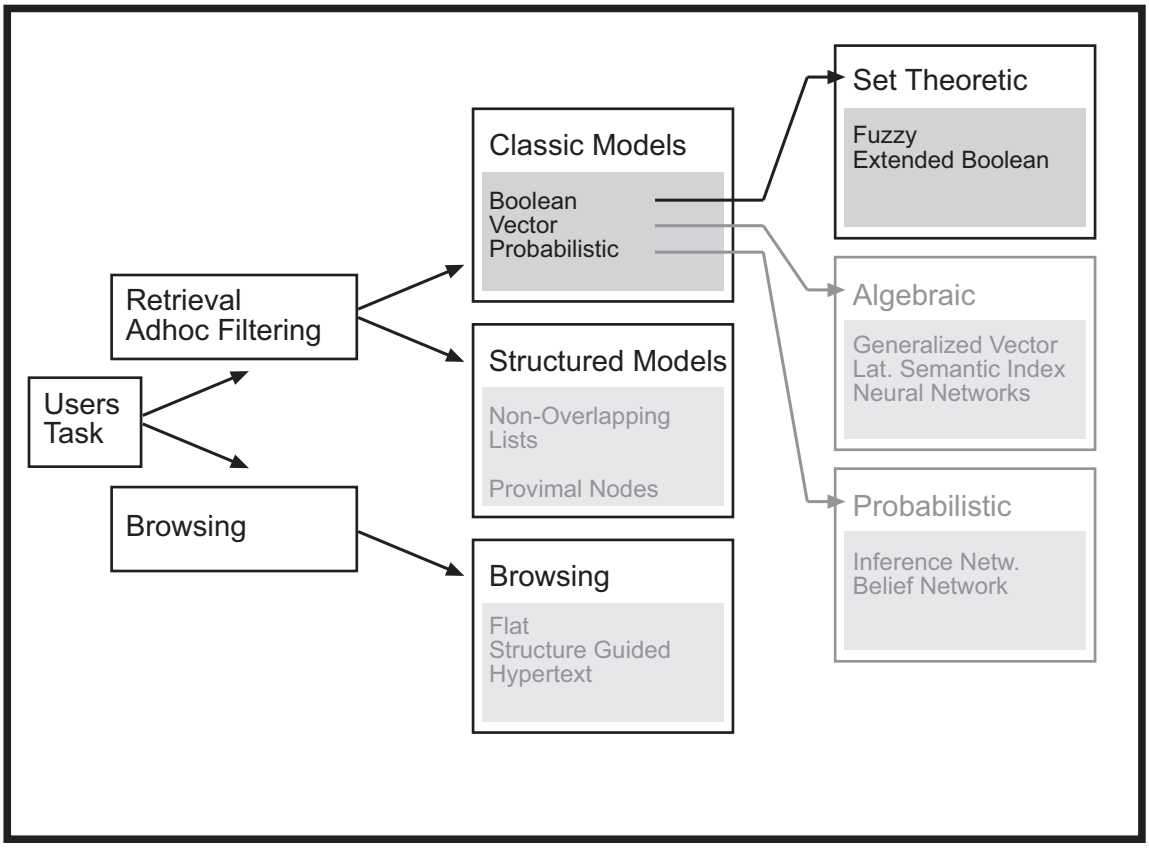
---

1 Baeza-Yates, R., Ribeiro-Neto, B. (1999), pp. 19-20

2 The light-colored models are named to provide completeness but won't be discussed in detail because that would go beyond the scope of this paper. Further information can be found in "Modern Information Retrieval" by Baeza-Yates and Ribeiro-Neto.

3 A more complex alternative is the dynamic creation of user profiles based on information about the preferences of the user. The profile adjusts to the preferences by user feedback. The user indicates not only documents considered to be of interest, but also the documents deemed non-relevant.

Figure 1:
Information Retrieval Models




Source: Baeza-Yates, R., Ribeiro-Neto, B. (1999), p. 21

In *classic information retrieval* each document of a collection is described by a set of keywords called index terms. Index terms are very likely to be nouns because their meaning, in most cases, is easy to identify. Adjectives and adverbs work mainly as complements and typically are less helpful. Index terms vary in their relevance. A word that occurs in every document of a collection is less useful than one found in only a few. Conversely, a word that occurs rarely within the collection has more potential to narrow down the number of documents returned.[4]

The oldest classic IR model is the *Boolean model*, in which index terms are related through *Boolean operators.*[5] Applying Boolean logic there are three possible types of Boolean search: AND search, OR search, and NOT search.[6] Boolean search is still used in commercial systems and can be quite powerful when the users are skilled in designing queries. In the Boolean model, a term is either present or absent in a document. Thus, the index term weights are binary

[4] Numerical weights of each index term are used to capture this effect. The numerical weight puts a figure on the importance of the index term when describing a document's semantic contents.

[5] Moens, M. (2000), p. 63

[6] Chowdhury, G.G., Chowdhury, S. (2001), p. 31

(0 or 1). There is no partial match of a query; each document is either relevant or non-relevant. This leads to the main disadvantage of the model - exact matching can lead to the retrieval of too many or too few documents. Index term weighting, employed in the vector model, can solve this problem.

The *vector model* eliminates the limitation of binary weights by computing a degree of similarity between each document of a collection. The usage of non-binary weights enables the vector model to enclose documents that only partially match query terms. This leads to a ranked set of documents matching the user's needs much more closely than a set retrieved using the Boolean model. To compute the ranking, the term weights are calculated using the concept of *clustering*: The goal of a simple cluster algorithm could be to separate a given collection of documents developed from a vague set description into two parts: one part with documents related to set, the other part with documents that are not. Use of the vector model results in a high retrieval performance, generating ranked answer sets difficult to improve upon. For this reason, it is currently a popular retrieval model.

The *probabilistic model* uses a framework based on statistical likelihood to solve the IR problem.[7] The query process seeks to find the ideal answer set containing the set of documents that is exactly relevant. After an initial estimate returning a first answer set according to the index terms, *user interaction* improves the probabilistic description of the ideal answer set. The user decides which documents are relevant and which are not. This information is used by the system to improve the description of the ideal answer set. It is expected, that the ideal answer description will develop and come close to the ideal answer set.

In IR, a *thesaurus* is used to apply *fuzzy set* theory,[8] defining relationships among terms and expanding the set of index terms to include related terms, thereby allowing for the retrieval of additional documents. A fuzzy set model typically employs a term-term correlation matrix, ranking documents relative to the user query.[9]

As discussed earlier, the Boolean model has the disadvantage of a lack of term weighting. The unranked answer set might be too large or too small to be useful. Because of this limitation, most new systems use some form of vector model. The *Extended Boolean Model* is an alternative to the vector model, extending its functionality with partial matching and term weighting. This allows for a combination of Boolean query formulation with the characteristics of the vector space model.[10]

---

7 Moens, M. (2000), pp. 64-65

8 John, R., Mooney, G. (2001), pp. 82-83

9 Baeza-Yates, R., Ribeiro-Neto, B. (1999), pp. 34-38

10 For example, in a query of two index terms connected with the operator “and”, the Boolean model treats documents containing only one term as irrelevant, just as it does documents containing neither. This binary decision criterion conflicts with common sense. The extended Boolean model can use different values for other operators (such as “and” or “not”). This allows for a combination of vector and Boolean search in one query.

*Structured Text Retrieval Models* combine information on text content with information about the structure of the document. For example, a user could search for a document containing the string 'air pollution' in italics and a figure labeled with the word 'earth'.

#### 2.1.2. Models for Browsing

A user is *browsing* if he explores a document by looking for interesting references. Users that browse are not interested in posing a specific query to the system. For both browsing and searching, the goal of the user is to find information. But in general, the searching task is clearer than the browsing task in the mind of the user.

The interactive navigational structure of hypertext allows for non-sequential text on a screen. This suggests a graph structure of nodes connected by directed links.

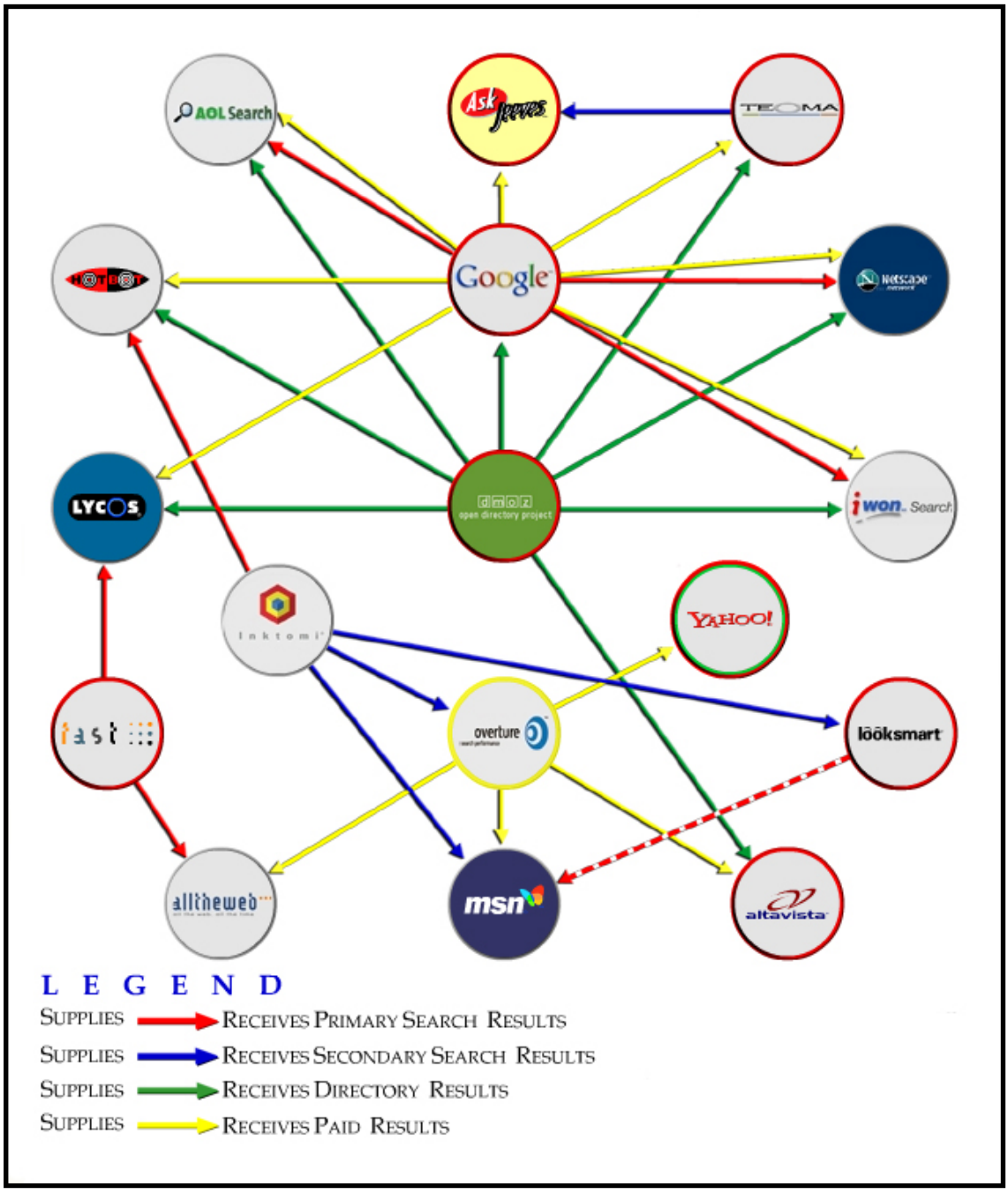
## 2.2. Web Searching

*Directories* and *crawler-based search engines* are the two major platforms used for searching the Web.[11] According to the user's task, there are less popular platforms for searching the Web that can be beneficial.

The following figures provide an overview of the current state of the search engine market and illustrate existing concentration tendencies.

[11] Appendix: Current Search Market

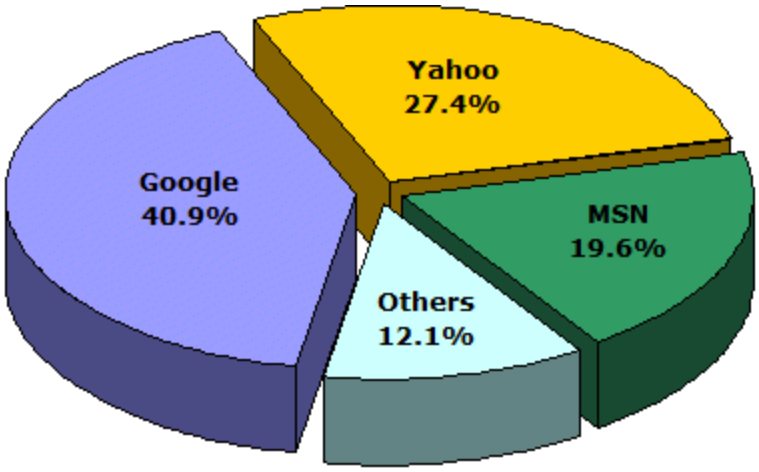
Figure 2:
Major Search Engines: Who Powers Whom?[12]




Source: Clay, B. (2004)

[12] An interactive version of this chart can be found online at: http://www.bruceclay.com/searchenginerelationshipchart.htm and on the attached CD-ROM.

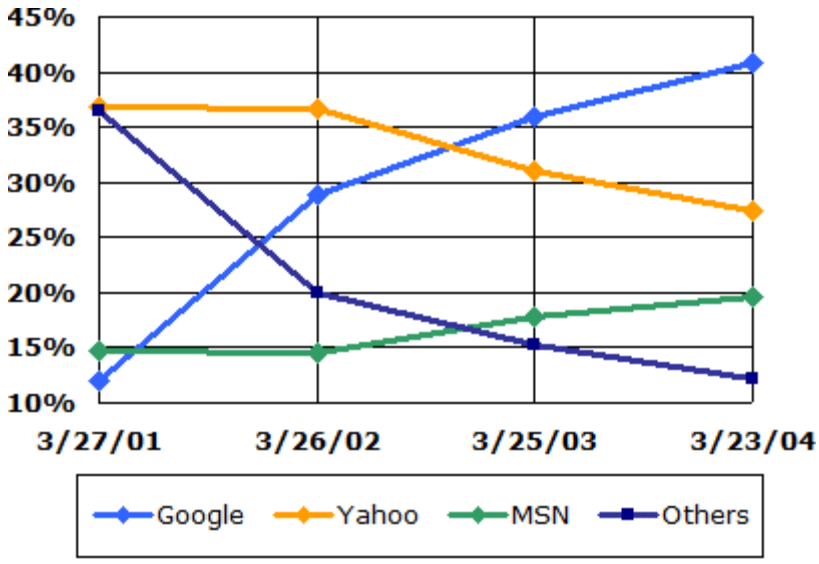
Figure 3:
Share of Search Referrals (March 2004)




Source: Sullivan, D. (2004d)

With 40.9 %, Google has the top share of referrals examined by WebSide-Story's StatMarket[13]. The following figure illustrates the recent success of Google. Google pulled traffic mainly form AltaVista and smaller services such as Exite or Infoseek/Go.[14]

Figure 4:
Referral Trends of Major Search Engines




Source: Sullivan, D. (2004d)

[13] http://www.statmarket.com/

[14] Sullivan, D. (2004d)

### 2.2.1. Directories

Directories are search engines powered by human beings assigning Internet resources to certain categories. Following the classification of the previous section, directories make use of the concept of browsing to find Web sites. Directories can be compared to 'Yellow Pages', where telephone numbers are sorted by branch and not alphabetically. In contrast to Yellow Pages, directories are not limited to companies; the user can browse through Internet resources classified into subject trees.[15]

Yahoo[16] was launched in 1994 and is the Web's oldest directory.[17] The Open Directory Project DMOZ[18] is the largest directory of the Web and is maintained by volunteer editors. It was taken over by Netscape (AOL) in 1998. The information from the directory can be used by anyone through an open licence agreement.[19]

### 2.2.2. Traditional Search Engines

Crawler-based search engines have become the most visited Web sites on the Internet, employing an information retrieval system to find relevant documents.[20] Search engines index sites automatically without the help of human editors, providing information about a much larger number of documents compared to directories.[21] Documents are retrieved according to query terms submitted by the user. The ranking of these documents depends on the algorithm used by the search engine.

Altavista[22] was the first crawler-based search engine on the Web. It started in 1995 and was the industry leader for several years, before being overtaken by Google.[23]

### 2.2.3. Niche Search Engines

Large search engines sometimes generate results that lack a sufficient degree of relevance. Niche search engines focus on a smaller range of topics, offering a more efficient and focused alternative to giants such as Google, often with en-

---

15 Babiak, U. (1999), pp. 52-55

16 http://www.yahoo.com

17 Sullvian, D. (2003)

18 http://www.dmoz.org

19 Ferber, R. (2003), pp. 299-300

20 Chang, G. et al. (2001), p. 4

21 Ferber, R. (2003), pp. 299-300

22 http://www.altavista.com

23 http://www.google.com

hanced features.[24] Topic-specific niche search engines are based on information filtering tools that gauge the relevance of crawled documents.[25]

CiteSeer[26] is a computer science research search engine with a number of unique capabilities, including citation indexing, links to related and similar documents, bibliographic coupling and collaborative filtering.[27] eBizSearch[28] is an experimental search engine based on CiteSeer's technology and focused on a very small niche: academic and commercially produced articles and reports about e-business.

### 2.2.4. Metasearch Engines

A metasearch engine sends a given query to multiple search engines, Web directories, and other data sources, then collects the results and formats them for display. Thus, the user can pose a query to many search engines through a single interface.[29] Metasearch engines differ from each other in the ranking of combined results and in the extent of translation of a user query into a specific query language. They are most effective when there is only a slight overlap in the indexes of the applied search engines and only a fraction of all Web sites is indexed by all search engines.[30]

Excite,[31] acquired by InfoSpace in 2002, was formerly a crawled-based search engine. It now uses the same underlying technology as the other InfoSpace metasearch engines, but maintains its own portal features. MetaCrawler[32] is one of the oldest metasearch services. It started in 1995 at the University of Washington and was purchased by InfoSpace in 1997.

### 2.2.5. Comparison Shopping Engines

These facilities are designed to help users find information about products sold online, such as product types, pricing, and online stores across the Web.

Froogle[33] is Google's entry, finding relevant products according to search terms.[34] The results are based on feeds from online stores and products. Yahoo! Shopping[35] is another example, connecting shoppers with thousands of

---

[24] Giles, C. et al. (2003), p. 413

[25] Chang, G. et al. (2001), pp. 17-18

[26] http://citeseer.nj.nec.com

[27] Giles, C. et al. (1998), p. 89

[28] http://www.ebizsearch.org

[29] Schwartz, C. (2001), pp. 112-114

[30] Baeza-Yates, R., Ribeiro-Neto, B. (1999), pp. 387-389

[31] http://www.exite.com

[32] http://www.metacrawler.com

[33] http://www.froogle.com

[34] Paulson, J. (2003), p. 1

[35] http://shopping.yahoo.com

merchants. But in contrast to crawler-based Froogle, its listings come from merchants hosted in Yahoo Store.

### 2.2.6. Web Question Answering Systems

Question-answering (QA) systems take the process a step further, actually synthesizing answers to queries. This 'deduction capability',[36] with its history in artificial intelligence (AI), differentiates them from other search engines. Their natural language interfaces are the trend for future information retrieval systems because they are more convenient for casual users.[37] QA systems provide users with the exact information needed, as opposed to a long list of documents that have to be looked through.[38] Direct answers to factual questions like 'What cities are within 100 kilometers of Cologne?' are provided by consulting a knowledge base.[39] This serves as a structured query interface for heterogeneous data from the multiple Web knowledge sources.[40]

START[41] is one of the first natural language QA systems with web interface. It was launched 1993 by MIT.[42] Ask.com is a commercial service with a natural question interface based on the work of hundreds of human editors mapping between question templates and sites. Wondir[43] is an example of a non-commercial information service providing a combination of live human answers, metasearch, and database searches. Its mission is to offer free information to anyone who asks, featuring a community of volunteers answering questions that cannot be handled by automated systems.[44]

---

[36] Zahdeh, L. (2003), p. 1

[37] Chang, G. et al. (2001), p. 9

[38] Lin, J. et al. (2003), p. 1

[39] Kwok, C. et al. (2001), pp. 242-244

[40] Katz, B. et al. (2002), pp. 1-2

[41] http://www.ai.mit.edu/projects/infolab

[42] Shah, U. et al. (2002), p. 462

[43] http://www.wondir.com

[44] Sherman, C. (2003)

# 3. Search Engines

Traditional IR models, as discussed previously, deal with information in well-structured databases. In contrast, the Web consists of a large mass of unstructured and unreliable document collections where the meaning of information is not always obvious.[45]

## 3.1. Challenges

The major challenge in search engine design is to construct a model that generates relevant and manageable results in response to well-formed queries. This is a complex task. Web search engines are faced with numerous problems not encountered in traditional information retrieval, where the focus is on relatively small, static, and homogeneous document collections. The Web contains an enormous amount of dynamic, heterogeneous, and hyperlinked information. This demands a technology able to rapidly search a vast amount of frequently updated documents. Storage space must be used efficiently and queries must be handled immediately. With the explosive growth of the Web, these tasks are becoming increasingly difficult.

### 3.1.1. Spam

Because search engines are commonly used to find information on the Internet, many Web developers pay careful attention to a site's ranking. Research indicates that 85 % of users tend to look only at the first result page.[46] This leads some site builders to attempt to manipulate their placement to get ranked within the top ten results.[47]

Search engines regard all techniques used to manipulate the relevance of a document as spam. Conversely, techniques used to make relevant documents more accessible or to prepare text for optimal usage are not regarded as spam.[48]

### 3.1.2. Content Quality

In traditional media, the publication of information is often associated with high costs and can therefore involve substantial financial risk. The Web is a new medium that enables publishers to operate at extremely low cost. In many cases there is no editorial process. As a result, information may be false, outdated, poorly written, or plagued with misspelled words, poor grammar, or OCR errors.[49] For a search engine, the challenge is to detect high-quality content for the retrieval.

---

[45] Machill, M. et al. (2003), pp. 18-19

[46] Silverstein, C. et al. (1998), p. 10

[47] Ferber, R. (2003), pp. 291-292

[48] Glöggler, M. (2003), p. 187

[49] Baeza-Yates, R., Ribeiro-Neto, B. (1999), p. 368; Nekrestyanov, I., Panteleeva, N.

### 3.1.3. Quality Evaluation

There are many new and promising ideas being developed for improving the relevance of search results with novel algorithms. For users and from a computer science perspective, it is important to determine if these new algorithms are effective.[50] Relevance is a concept that is intuitively understood, but very tricky to define.[51] Aspects of retrieval efficiency, e.g., speed or processing, can be measured precisely. Aspects of retrieval effectiveness, e.g., the ability of a system to satisfy the information needs of a user, are more complex and traditionally involve IR laboratory experiments.[52] One aspect of effectiveness is *precision*, the relevance of a return to the user. Another important element is *recall*, the amount of relevant information made available to the user.

### 3.1.4. Invisible Web

Only a portion of the Web can be accessed through the use of search engines. Content not searchable is sometimes called the 'invisible Web' or the 'deep Web' and exists in specialized databases.[53] A document's inaccessibility can result from various factors.[54] Most spiders ignore dynamic Web pages that are based on databases and exclude dynamic documents that contain a '?' in the URL. Additionally, many pages are protected with passwords. A search engine may, in effect, show the front door of a library but not its content.[55] The content can only be searched using the retrieval tools offered by the database.

Because a large proportion of the invisible Web is found in topic-centered databases produced by professional content providers, it is often of a higher quality than that of other Web documents.[56] It is therefore very important for search engines to find a way to index its content to obtain an acceptable level of quality. Cooperation with content providers is important in this regard.

## 3.2. Search Engine Architecture

Figure 2 shows the basic architecture of an Internet search engine: a *crawler*, an *indexer*, and a *query engine*.[57]

The *crawler* starts from a given set of URLs and follows links to reach other pages. All URLs appearing in the retrieved documents are parsed to the crawler control module, which determines the links to be visited next by the crawler. Re-

---

(2002), p. 208

50 Hawking, D. et al. (1999), pp. 243-244

51 Ingwersen, P. (2000), pp. 167-171

52 Robertson, S. (2000), pp. 81-82; Robertson, S. (2002), pp. 258-260

53 McGuigan, G. (2003), pp. 68-69

54 Baeza-Yates, R. (2003), p. 99

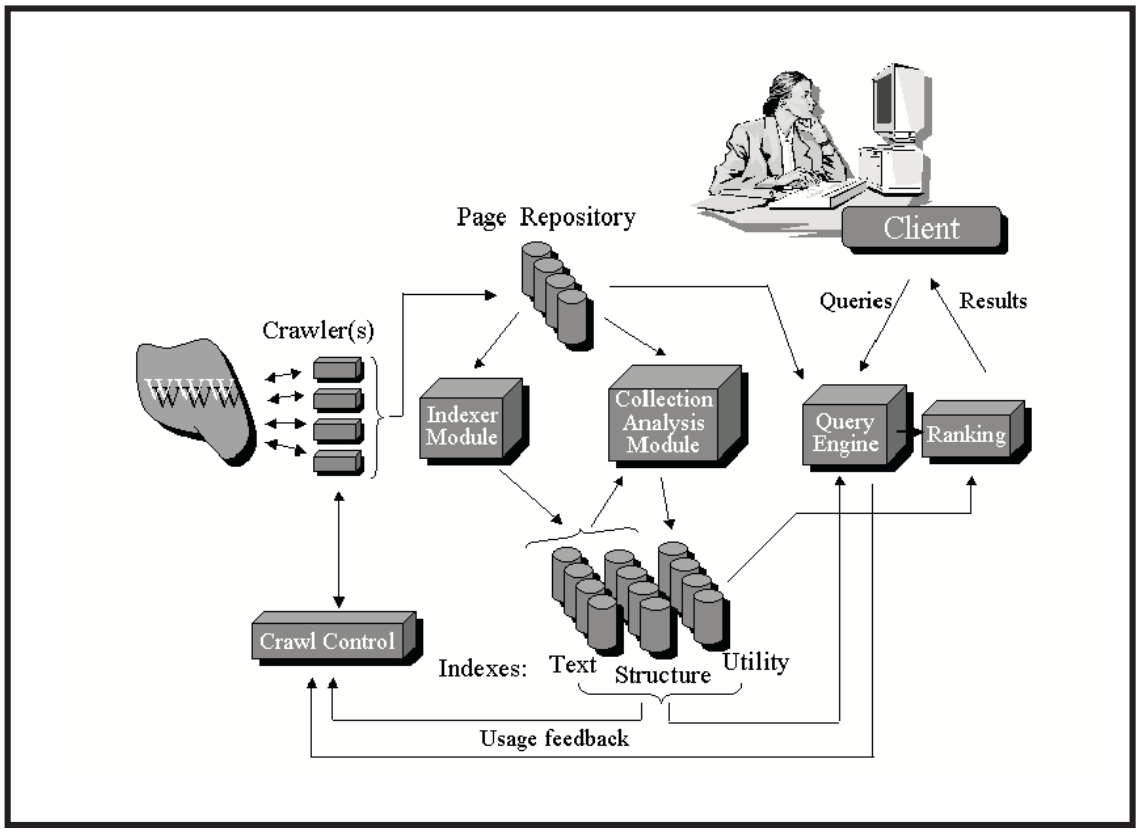
55 Ludwig, M. (2003), p. 8

56 Machill, M. et al. (2003), pp. 43-44

57 Arasu, A. et al. (2001), pp. 2-30; Baeza-Yates, R., Ribeiro-Neto, B. (1999), pp. 373-374

trieved documents are stored in the page repository. The *indexer* module extracts all the words in the documents and assigns the URLs where each word was found. The result is a huge 'lookup table' that shows all URLs that point to pages containing a certain word. The *query engine* module compiles search results that depend on the indexes. There is typically a very large set of result documents, then sorted for relevance by the ranking module.

Figure 5:
Search Engine Architecture




Source: Arasu, A. et al. (2001)

### 3.2.1. Crawler Module

The crawler module downloads Web pages for later processing in the indexing module. The crawler starts at a particular set of URLs and then accesses all the links from it.[58] Newly found URLs are put in a queue to be visited and scanned by the crawler. In this manner, it traverses the graph formed by the Web

Since crawlers cannot download all documents, they prioritize the URLs and visit the 'important' Web pages first. This potentially difficult process is known as *page selection*. *Page refresh* is another problem, given that Web pages are updated at very different rates. Thus, the crawler has to decide whether or not to revisit a site. This has a significant impact on 'freshness'. For example, a news site will likely change its content more often than many other sites.

[58] Bradman, O. et al. (2000), p. 9

The *page repository* is a system that stores large collections of Web pages. It provides an interface for the crawlers to store downloaded Web pages that can then be accessed efficiently by the indexer and the collection analysis modules.

### 3.2.2. Indexer

The indexer reads the text of documents, building both a text index and a link index. The *collection analysis* module builds a range of further useful indexes based on these two indexes.

The inverted indexes are perhaps the most important structure for text-based retrieval (*text index*). They consist of a set of inverted lists for each word. An inverted list provides information about the appearance of certain words in a collection in the form of a sorted location list. In a simple case, location information consists of a page identifier and the position of a word in that page. Search engines can use other information about a word's occurrence and store it in additional payload fields. For example, words in headings, boldface, italics, or anchor text might be weighted more heavily in the computation of a ranking. Most *text indexes* also contain a lexicon, a list of all terms occurring in the index.

The index of a document's link structure (*link index*) is built by modeling the Web as a graph with nodes and edges. Each page represents a node; the hyperlink between two pages is an edge. The inverted Web graph provides neighborhood information such as the set of pages pointing to a given page. It is used by link-based ranking algorithms like PageRank or HITS for calculating a page's importance. Neighborhood information is also employed in the retrieval of 'related' pages.

The collection analysis module is used to build *utility indexes*. A search engine's features and its ranking algorithm determine the number and style of these indexes. An index listing all the pages on a site is required for a search engine that allows searches restricted to that site. This index shows a list of all pages that belong to a certain domain. Another example is the PageRank algorithm that needs pre-computed neighborhood information about the importance of a page for the ranking at query time.

### 3.2.3. Document Preprocessing

Document preprocessing specifies the words used as index terms. Not all words in a document are of the same importance semantically. Nouns are considered most representative because they tend to carry greater meaning. In contrast, a word like 'the' is rarely important and its use might lead to the retrieval of irrelevant documents. Five operations characterize these distinctions, controlling the size of the vocabulary, reducing noise, and improving retrieval performance.[59]

[59] Baeza-Yates, R., Ribeiro-Neto, B. (1999), pp. 163-190

*Lexical Analysis* involves the conversion of a stream of characters into a stream of words.[60] The recognition of words may seem like a simple task, but there are some cases, that require careful consideration. Abbreviations (e.g., 'cf.'), potentially confused with the last word of a sentence, need to be identified. Hyphenated words represent another difficult context-dependent decision. Due to inconsistent inclusion of hyphens, a break-up of hyphenated words might be useful. This can be problematic, however. The hyphen may be part of a proper name (e.g., 'MS-DOS') or be used to split a single word at the end of a line into two syllables. Numbers are usually not selected as index terms; without surrounding context, their meaning is typically indeterminate. On the other hand, a 16-digit number that identifies a credit card could be highly relevant.

*Stopwords* are words that occur in nearly all documents of a collection. They are not good discriminators and generally should not be chosen as index terms. Articles and prepositions are commonly included in a so-called stoplist, useful in reducing the size of the index structure. This compression can provide a higher performance but can also reduce recall. For example after elimination of stopwords, the phrase "to be or not to be" might leave only the term 'be'.

*Stemming* is a process of substituting words with their respective stems to improve the relevant match between the terms of a query and a page's text. Plural or past tense suffixes are examples of variations that could prevent an exact match between query and document content. For instance, the stem of the word 'connect' has variants like 'connected', 'connecting', 'connection', and 'connections'. Stemming increases retrieval performance by eliminating variants of the same root word, thereby reducing the size of the indexing structure.

In a full text representation, all the words in a document are used as index terms. Otherwise an *index term selection* is needed to identify the index terms. In an automatic approach, index terms are selected. The identification of noun groups, chosen for their semantic weight, is a common practice.

In a simple case, the *thesaurus* consists of a precompiled list of words important in a given domain of knowledge and a set of related, often synonymic, words for each word in that list. More complex thesauri involve the normalization of vocabulary and the inclusion of phrases.[61]

### 3.2.4. Query Module

The query engine module receives and fills search requests from users, relying heavily on the indexes and sometimes on the page repository. Due to the Web's size and because users typically enter only one or two keywords, result sets are often very large.

---

[60] Moens, M. (2000), pp. 78-80

[61] Baeza-Yates, R., Ribeiro-Neto, B. (1999), pp. 170

### 3.2.5. Ranking Module

The ranking module sorts the results a return by their relevance to a query. Most traditional techniques rely on measuring the similarity of query texts with texts in a collection's documents.

## 3.3. Ranking Techniques

A key to successful search engine design lies in the development of effective algorithms.[62] Unfortunately, many of the techniques used in commercial search engines are not published in detail.[63] Most search engines use a variation of Boolean and vector model for ranking.[64] IR should involve not only the content of documents, but also the logical structure, the layout, and external attributes of documents, as well as user involvement.[65]

### 3.3.1. Internal Content

When internal content is used for ranking, documents are rated highly if the query terms occur frequently. Visual presentation details, such as the font size of words, can also be relevant for ranking: Words formatted in a larger or bolder font are typically weighted more heavily than other words.

The *Term Frequency (TF) Algorithm* is based on Zipf's law. As a word's occurrence in a document increases, its relevance rises.[66] The simplest form is the counting of keywords, where term frequency measures the occurrence of a keyword in a document. This method is not suited for short texts because most terms likely occur only once or twice.

The *Inverse Document Frequency (IDF) Algorithm* is collection-dependent and weights a keyword in relation to its frequency.[67] Words that occur in numerous documents of a collection generally make poor indicators. For example, the term 'computer' is not suitable as an index term in a computing collection if it occurs in nearly every document. On the other hand, a word appearing in only a few documents of a large collection is useful because it narrows down significantly the number of documents in a return. The IDF Algorithm assigns weight to a query term, where low frequency terms are likely to point to relevant documents. An improvement is the *TF.IDF weighting*, where term frequency weights are multiplied by collection frequency weights.[68]

Information about the *logical structure* of a document can also be used to weight query terms. In this approach, a term that occurs in the title of a docu-

---

[62] Kobayashi, M., Takeda, K. (2000), p. 146

[63] Arasu, A. et al. (2001), p. 6

[64] Baeza-Yates, R., Ribeiro-Neto, B. (1999), p. 380

[65] Fuhr, N. (2000), p. 49

[66] Glöggler, M. (2003), pp. 76-77

[67] Moens, M. (2000), pp. 91-92

[68] Moldovan, D., Surdeanu, M. (2003), p. 130

ment would be regarded as more relevant than one found in the body.[69] The relative position of a term can also be used for weighting. A term that occurs at the beginning of a page might reasonably be weighted more heavily than a term found in the middle or end of a document. Announcements from news agencies, for example, generally start with the important facts and follow with background information. For this reason, some search engines weight terms that occur in the beginning of a text more heavily.

### 3.3.2. Usage Information

The technology of *click popularity* was developed by DirectHit.[70] It collects information on the queries individual users submit to search services, the pages they look at subsequently, and the time spent on each page. This information is used to return pages that most users visit after deploying the given query. Thus, the ranking depends on user activity.

Alexa, a search engine of Amazon, uses another approach to exploit usage information. Alexa observes the movements of registered users to find typical surf patterns to sites with similar content.[71] Alexa[72] partnered in 2002 with Google and now provides additional information to its search result.

### 3.3.3. Link-Based

This approach involves analyzing the hyperlinks between Web pages. Vast scale and highly variable content quality can give poor results in traditional information retrieval techniques. Web search results can be improved by using the information provided by link structure between the pages. Hypertext Induced Topic Search (HITS) and Google's PageRank are the best-known of these algorithms.[73] The Web is modeled as a graph containing a node for each page and a directed edge between linked pages. This link graph can be used for ranking, finding related pages, and various other purposes.[74] The methods of link analysis are based on the following assumptions:[75]

An author recommends another page by setting a link to it. Pages with many incoming links are highly recommended.[76] This can be implemented by a simple link count or a calculation of page weight, where links from popular sites are held to be of greater value. Linked pages are more likely to be about the same

---

[69] Ferber, R. (2003), p. 71

[70] http:// www.directhit.com

[71] Machill, M. et al. (2002), p. 30

[72] http://www.alexa.com/

[73] Richardson, M., Domingos, P. (2002), p. 1441; Wensi, X. et al. (2002), p. 146; Agosti, M., Melucci, M. (2000), pp. 259-269

[74] Henzinger, M. (2000), p. 3

[75] Craswell, N. et al. (2001), pp. 250-251

[76] Chang, G. et al. (2001), p. 101 ; Walker, J. (2002), p. 72

topic than those that are not linked.[77] The anchor text of a link describes the content of the target page.

The underlying principle is that a Web page referenced by many sites is likely to be more important than a page that is rarely referenced. This is analogous to citation analysis, where an article widely cited is considered better than one infrequently cited.[78]

Search engines can associate the text of a link with the page that the link is on and in addition with the page the link points to. The use of *pointing anchors* has advantages. First, anchors often give a better description of Web pages than the pages themselves. Second, anchors may exist for document elements that cannot be indexed by a text-based search engine, such as images. This makes it possible to return information that has not been crawled.

A simple *query-independent* method can score pages according to the number of links pointing to the page: the larger the number of hyperlinks, the better the page. In this approach, every link has the same importance. There is no difference between links from high- and low-quality pages. This simplified ranking is not a very effective measure, since it is quite easy to artificially create a lot of pages to point to a certain page through spamming.[79]

*PageRank* is a solution to this problem of manipulation. Developed by Brin and Page at Stanford University, it is the technique underlying Google's global ranking scheme.[80] PageRank is designed to capture the importance of a page, extending the basic citation idea by measuring the importance of documents that point to a given page. This calculation of global importance is called rank value.[81] Google's success demonstrates that PageRank is an effective method of ranking pages. However, as Beaza-Yates et al showed, PageRank tends to give higher ranks for older pages. They propose a modification of PageRank to correct this bias against newer pages.[82] Pretto proposes a personalized PageRank depening upon user feedback regarding importance.[83]

The *Hypertext Induced Topic Search (HITS)* algorithm was first proposed by Kleinberg.[84] It is a *query-dependent* ranking method that creates scores for authority and hub sites. Authority pages are those most likely to be relevant to a particular query because many links point to them.[85] Hub pages are not necessarily authorities themselves but at least point to several authority pages. This creates a positive two-way feedback in which better authority pages are linked

---

77 Richardson, M., Domingos, P. (2003), p. 16

78 Chau, M., Chen, H. (2003), pp. 199-200

79 Henzinger, M. (2000a), p. 4

80 Brin, S., Page, L. (1998), pp. 3-4

81 Chen, Y. et al. (2002), p. 550

82 Baeza-Yates, R. et al. (2002), pp. 117-126

83 Pretto, L. (2002), pp. 138-143

84 Kleinberg, J. (1999), pp. 604-630

85 Henzinger, M. (2000), p. 5; pp. Diligenti, M., et al. (2002), 511-512

from good hubs and better hub pages come from links of good authorities.[86] Topic distillation[87] describes the process of finding high-quality Web sites according to a query topic. It seeks to use only the most authoritative sources instead of all documents considered relevant.

Smider[88] is a search engine inspired by Kleinberg's HITS algorithm. It identifies topic-related communities crawled according to the query topic.

## 3.4. Trends

The ideal search engine is a tool that understands any query, no matter how poorly constructed, and immediately retrieves the perfect resource.[89] No single innovation is sufficient to achieve this goal. The technologies presented in this section illustrate trends in search engine development.

### 3.4.1. Usability, Visualization, and User Interface

The easy interpretation of a mental desire of information into a useful query is one of the greatest challenges in search engine design. Most users do not want to learn a search language that would enable them to skillfully restrict a search and retrieve relevant documents.[90] Vague queries with one or two words are common.

Google's success indicates that many users do not want a broad personalized portal with a lot of confusing features. They simply want quick and easy access to relevant information. Usability can be improved by anticipating faulty operation. When integrating intelligent features (like the correction of spelling errors), it is important to provide the impression of a utility that is easy to use.

Methods of intuitive user guidance try to specify the query of the user.[91] Vivisimo[92] is an example of a metasearch engine that automatically clusters search results into categories selected from words and phrases contained in the search results themselves. If the Internet is seen as a giant facility in which books are spread on the floor, Vivisimo could be seen as a librarian who places them on shelves in a way that makes sense.[93] This technique allows for effective queries of only one or two words, reducing the number of hits to a manageable amount. A spell check function is another example of cluster analysis.[94] In contrast to

---

[86] Baeza-Yates, R. et al. (2002), pp. 380-381

[87] Schimkat, R. et al. (2002), pp. 8-9

[88] http://frank.spieleck.de/metasuch/

[89] Kline, V. (2002), p. 252

[90] Eastman, C., Jansen, B. (2003), p. 384

[91] Machill, M. (2003), pp. 39-41

[92] http://www.vivisimo.com

[93] Bergstein, B. (2004)

[94] Document clustering is the automatic organization of documents into groups or clusters with no further human intervention required.

conventional systems that use dictionaries, new systems use past queries to deliver the correct formation.

Most search engine designs present search results as ranked lists of documents with information like URL, title, and keywords in the return.[95] Methods of visualization allow the presentation of hits according to their context,[96] where maps enable the user to see the context of documents and intuitively expand the query. KartOO[97] and Mooter[98] are examples of metasearch engines that offer search results employing clustering.[99]

### 3.4.2. Semantic Web

As the Web expands with sites written in natural language, finding specific information becomes increasingly difficult.[100] Berners-Lee, the inventor of the WWW, envisioned a Semantic Web capturing the meaning of content.[101] The idea is that the Web as a whole can be made more intelligent and intuitive about serving user needs with the help of machine-understandable semantics.

Although search engines index a large amount of the Web's content, they are limited in their ability to show the exact pages a user wants. The Semantic Web would bring structure to content by giving a defined meaning to information. The enrichment of content with formal semantics enables search engines to reason about the content of Web sites. "Expressing meaning" is the main task of the Semantic Web.[102]

Berners-Lee predicted a five-layer architecture in which developers and authors use self-descriptions and other techniques so programs can interpret context and selectively find what users want. With the use of background knowledge like the semantic correspondence between words, they are more likely to return relevant information. The Semantic Web uses additional sources to add information about meaning with the help of ontologies.[103]

### 3.4.3. Comparison Shopping Enhancements

Buyers compare products not only by price, but also by several other conditions, like discounts or return policy. It is difficult to develop an intelligent system that supports matchmaking of current E-commerce sites because these complex product conditions are typically described inconsistently in natural lan-

---

[95] Kobayashi, M., Takeda, K. (2000), p. 157

[96] Machill, M. et al. (2003), pp.41-42

[97] http://www.karoo.com

[98] http://www.mooter.com

[99] Dvorak, J. C. (2004)

[100] Choi, O. et al. (2003), p. 588

[101] Berners-Lee, T. et al. (2001), Shah, U. et al. (2002), pp. 461-462

[102] Bozsak, E. et al. (2002), p. 305

[103] Stuckenschmidt, H. (2002), pp. 114-115, Doan, A. et al. (2002), pp.1-3

guage. The standardization project RuleML tries to solve this problem by defining a set of rules.

Customers are sometimes provided incentives, special conditions designed to encourage the purchase of a subclass of products. For example, most theme parks offer weekday discounts. Airline ticket pricing often includes details regarding advance purchase, minimum stay, and stopovers. These incentives, established according to a vendor's intentions, often result in complex price structures, making evaluations of available offers a difficult task. Moreover, the structure of incentives between vendors can vary widely and they usually do not correspond with the motivations of buyers.[104]

### 3.4.4. Query Reformulation and Cross-Language Retrieval

While effective use of Web search engines requires careful construction of search queries, additional relevant documents can be retrieved by a process of assisted reformulation.[105] This involves first expanding the query term with additional terms from relevant documents, then reweighting the terms of the expanded query.[106]

A popular reformulation strategy is *relevance feedback*, where the user examines retrieved documents and indicates which documents are relevant and which are not. Important terms and expressions are given greater weight in the reformulation of the query, leading to the retrieval of more relevant documents. This feedback cycle provides additional cluster support, built through user interactivity. The identification of terms related to the query term can automatically provide the description of a larger cluster of relevant documents.

Many Web pages are published in languages other than English and many Internet users are non-native English speakers that can read and understand English but feel uncomfortable in formulating queries in a foreign language.[107] Even if the user knows several languages well, there is still the problem, that the query must be formulated in each language. A query translation tool would be useful.

A simple dictionary based translation tool could be extended with relevance feedback and thesaurus-based increase of keywords.[108] Another method to realize the translation is the usage of large parallel texts that are found on multilingual Web sites. Existing search engines could be used to find Web pages on the net to build the needed parallel corpora automatically.[109]

---

[104] Kozawa, M. (2002), pp. 152-154

[105] Leroy, G. et al. (2003), pp. 230-231

[106] Baeza-Yates, R., Ribeiro-Neto, B. (1999), pp. 117-139

[107] Cross-Language IR involves querying multilingual collections in one language to retrieve documents in other languages, and belongs to the domain of Multilingual Information Access. See Peters, C., Sheridan, P. (2000), p. 52

[108] Sadat, F. et al. (2002), p. 114

[109] Nie, J., Chen, J. (2003), pp. 218-219

### 3.4.5. Hybrid Search

Many content providers charge for high-quality content hidden in the invisible Web; search engine designers seek means to index this content.[110] Scirus,[111] operated by Reed Elsevier, is an example of a hybrid search engine that includes paid content. Beyond returns from scientific Web sites, it points to documents other search engines don't index. This allows access-controlled sites like scientific journals to be searched. A hybrid search engine requires cooperation with content providers and contracts between all parties.[112]

### 3.4.6. Natural Language Processing and Web Question Answering

In field of natural language processing (NLP), search engine designers work with linguists to understand the meaning of text and act accordingly. NLP-based search engines use algorithms that rely on phonetics, morphology, syntax, semantics, discourse, and pragmatic content. For instance, when searching on Europe, a return might be improved by including France, Germany, and Italy in the query.[113]

In contrast to Information Extraction (IE) systems, where templates are filled with information from a defined domain of extraction, IR systems do not specify exactly where answers are located. Question Answering (QA) systems formulate answers to natural language questions, taking IR and IE a step further.[114] Web-based QA systems provide an intuitive user interface for searching, with answers reported in natural language instead of a large set of documents.

In a generic QA system, search engines could replace the static collection of documents. This does, however, create a few potential problems. First, the QA system must understand the syntax of the search engines. For instance, some search engines differ in their use of Boolean operators. Second, network latency and the problem of inaccessible documents should be considered. Finally, inconsistent Web document structures must be transformed.

### 3.4.7. Personalized Search

The relevancy of search results can be improved significantly through *personalized search*,[115] where user interest is used to refine a return.[116] Applications of this feature have been limited by privacy concerns and technical problems. Users fear that a detailed profile of searched terms and visited sites might be

---

[110] If the market accepts paid content, search engines must be able to index it. Otherwise, they lose value as service providers.

[111] http://www.scirus.com

[112] Machill, M. et al. (2003), pp. 43-45

[113] Rappoport, A. (2000), pp. 12-13

[114] Moldovan, D., Surdeanu, M. (2003), pp. 129-145

[115] Jeh, G., Widom, J. (2003), p. 271

[116] E.g., a teenager searching for literature might have interests different from those of a senior citizen.

made available to marketers or government institutions.[117] The distinction between techniques that can enhance user experience (like Google's cookie) and a registration process that collects personal data such as real name, age, and address (like Yahoo membership) is important in this context.[118]

Google recently started a beta test for Gmail[119], a free webmail account with 1000 MB of free storage that promises to allow users to search their email as easily as the web, providing the company a closer connection to the user.[120] This service is financed with contextual advertisements, enabling the construction of user profiles that could be important for further developments in personalized search. Privacy concerns are raised by the fact that Google scans the content of private mail to include relevant ads.[121] Organizations like the World Privacy Forum have a view of data gathering very different from that of many corporations. In an open letter,[122] they have expressed a fear that companies and even governments will walk through the email-scanning door once it is made available. If people come to accept having the content of private email scanned for ad delivery, exploitation could become a real possibility.

Google's experimental Lab[123] has presented a beta version of a search engine that tailors search results according to user preferences and a news service that alerts users when new information on a specific topic is found.[124] The search is personalized manually by the user through categories. With the help of a slider, results can be rearranged to add more or less emphasis to profile information. Future systems should be able to adapt automatically to preferences and give users a chance to define different roles for themselves.

Yahoo[125] allows users to personalize content with modules.[126] A user can change layout and pick content modules that are of interest (e.g., a document's news or travel information). The content of the modules is filtered according to preferences the user has submitted. Some content, like sports events, is customized automatically according the zip code of the user. The search can be further customized for language, the number of documents to be retrieved, and the use of an adult filter.

---

[117] Thompson, B. (2003)

[118] Sullivan, D. (2003a)

[119] http://gmail.google.com/

[120] Sullivan, D. (2004)

[121] Markoff, J. (2004)

[122] http://www.worldprivacyforum.org/gmailopenletter.pdf

[123] http://labs.google.com/

[124] Parker, P. (2004)

[125] http://my.yahoo.com/

[126] Rossi, G. et al. (2001), p. 276

### 3.4.8. Local Search

The search for local information is often frustrating; isolating relevant data can be difficult. Users who want information on a local restaurant or hairdresser are not interested in results from other parts of the world. Consequently, *local search* has recently become the focus of aggressive attention and effort of all major search engines.[127] It has the potential to combine the advantages of the profitable Yellow Pages industry and advertising revenue from small and medium-sized businesses with the convenient interface of a keyword-based search engine. The challenge for designers is to find ways to filter irrelevant or unstructured data.[128]

A recent study of information about New Zealand found no significant benefit using a local search engine.[129] One possible explanation for this outcome is that the engine lacked adequate features. Another reason may be that there are a lot of relevant documents outside local .nz domains.

Nevertheless, there is a big potential for local search engines. Citysearch[130] is an example of a local search service providing information about businesses like restaurants, retail merchants, travel agents, and professional services. Google has recently added structured data from Yellow Pages to the beta version of Google Local.[131]

### 3.4.9. Social Networks

A *social network* is established by all the relationships an individual has with family, colleagues, neighbors membership groups, etc.[132] On the Internet, *social network sites* make connections between individuals based on recommendations from friends.[133] Friendster[134] was one of the first social network communities, connecting people though networks of friends and friends-of-friends.

Eurekster[135] is an example for a search engine that combines the concept of social networks with traditional search engine technology. In contrast to the common practice of personalized search, results in this context do not depend on who you are, but rather on whom you know.[136] In this approach, a personalized search allows users to share popular Web sites within a community.[137]

---

[127] Sterling, G. (2004)

[128] The unstructured content of Web pages has to be enriched with more structured data like that found in a Yellow Pages format.

[129] Smith, A. (2003), pp. 104-107

[130] http://www.citysearch.com

[131] http://local.google.com

[132] Pujol, J. et al. (2003), p. 381; Zhong, N. (2003), pp. 8-9

[133] Reardon, M. (2004)

[134] http://www.friendster.com/

[135] http://www.eurekster.com/

[136] Sullivan, D. (2004a)

[137] In contrast to traditional user feedback, this technique is much more resistant to spam because a manipulation would only affect a ranking within the network.

Users invite friends to join a social network and preferred results of network members appear at the top of the return.[138] Google's release of a social networking service called Orkut[139] demonstrates the popularity of this concept.[140]

### 3.5. Revenue Sources

Google's planned IPO is expected to be the biggest since the dot-com bubble burst, possibly reaching a valuation of as much as $25 billion USD.[141] TNS Media Intelligence/CMR ranks Yahoo third among Internet companies in advertising revenues ($807 million in 2002, an increase of more than 35 % from the previous year).[142] The company's executives estimate that each percent of the search market share is worth $200 million in revenues generated from paid listings.[143]

These examples illustrate the expanding market for search engines and their potential as a business model. Most users do not want to pay for search services. Thus, entrepreneurs in this market must find other sources of revenue such as advertising clients, content providers, portals, and even other search engines. The following figure provides a classification of possible sources. It is important to seek an optimal combination.

Figure 6:
Revenue Sources

Revenue Sources

| Direct Revenues from Users | | Indirect Revenues |
|---|---|---|
| Dependent of Usage | Independent of Usage | |
| **Hybrid Search Engines**<br><br>**Paid Search**<br>- depending on volume<br>- depending on time | **Subscription**<br><br>**Subvention** | **Advertising**<br>- Banner<br>- Paid Listing<br>- Contextual Ads<br><br>**Paid Inclusion / Paid Submission**<br><br>**Data Mining**<br><br>**Licenced Technology**<br>- Infrastructure<br>- Search Results |

Source: own representation

[138] Gaither, C. (2004)

[139] http://www.orkut.com/

[140] Sullivan, D. (2004b)

[141] Salkever, A. (2004)

[142] Endicott, R. C. et al. (2004), p. 38

[143] Morrissey, B. (2004)

### 3.5.1. Direct Revenues from Users

Some users are willing to pay for high-quality content not affected by the interests of third parties. Scientific researchers, for example, rely on respected journals and other credible sources. Some hybrid search engines include this type of paid content in their results.

Among traditional media, the pay TV market has demonstrated that some consumers prefer high-quality programming not interrupted by advertising. Search providers could adopt this strategy by charging directly for services or offering subscriptions.

AllAcademic[144] is an example of a subscription-based niche search engine for research needs presenting peer-reviewed conference proceedings and journal articles from the social sciences. Users pay for this service because they get an interface for access to high-quality, reliable content hidden in the deep Web.[145]

### 3.5.2. Indirect Revenues

The lion's share of search engine revenue is generated through advertising (paid listings), which allows unpaid editorial listings to be provided for free.[146] This same pattern is found among traditional media. But in contrast to promotions in newspaper and television outlets, paid and editorial listings on the Web can easily be mixed. Since this is widely recognized, search engines working to build reputation and long-term relationships with satisfied users are well-advised to label their paid listings as such.

Some search engines include *banner ads*[147] in result pages. Advertising clients typically contract for the number of times the banner was included.[148] *Keyword banners* appear for specified words entered in a query as a means to narrow down the target group; random banners appear by chance.[149]

While traditional search engines rank documents by relevance, *paid placement* services rank them in relation to the fee paid for a chosen keyword.[150] Nearly every major search engine with significant traffic accepts paid listings. This unique form of advertising means that the owner (or stakeholder) of a Web site

---

[144] http://www.allacademic.com/explore.html

[145] Dvorak, J. C. (2004)

[146] Sullivan, D. (2003b)

[147] A banner is a graphic display linked to an advertiser's Web page that contains a short text or graphic message to promote a product.

[148] Advertising clients can be charged in a variety of ways: first, by *view*, where payment is based on the number of times an ad is displayed; second, by *click*, where the client pays only for the number of visitors directed to the site; third, by *lead*, where payment is collected only if the directed visitor fulfils a prescribed lead such as registering for a special service; and finally, by *sale*, where a commission is charged for visitors who actually buy something.

[149] Turban, E. et al. (2000), pp. 119-124

[150] Goh, D., Ang, R. (2003), p. 87

can be guaranteed his pages will appear in the top results for specified terms. Clients bid for keywords that describe a site and pay for every click the search engine sends them. Overture[151] is the oldest paid placement search engine, and perhaps the most important because it distributes its listings to a wide range of major service providers, including AltaVista, AOL Search, Lycos, HotBot, and Netscape Search.

Google runs a program called AdWords[152] that places paid listings on its own site as well as on some others. Clients bid for placement and pay each time someone clicks through. AdWords included on other sites (in a programme called AdSense.[153]) are known as *contextual ads*. AdSense is designed for publishers who want to receive revenues from text-based ads relevant to the content of their pages.

Contextual advertising contrasts with the pay-for-performance model in that results do not appear on the search engine, but rather on pages of other sites.[154] Kandoodle[155] is a search engine that offers only contextual marketing, where clients can select a specific category for their ads.[156]

A content provider can pay to have a site included in the index of a search engine. These *paid inclusion* programms guarantee placement but do not address ranking. There is no guarantee that included pages rank well, but sites enrolled in paid inclusion programms are likely to receive more visitors than those that are not.[157] In contrast to paid inclusion, a *paid submission* service charges for the application of a Web site but there is no guarantee that it will be accepted.

Detailed data about consumers are a valuable resource that search engines are able to generate.[158] Through *data mining*, the search activity of a user can enrich a profile for direct marketing. Yahoo announced a program in which the search activities of their e-mail users would be tracked and used to promote search-related products of third party vendors. The service, called Yahoo Impulse Mail, would affect only those who opt in.[159] Users who want to keep their searches private would simply log out as a Yahoo member.[160] Perhaps because of unfavorable public reaction, the service has not yet started.

---

[151] http://www.overture.com

[152] http://adwords.google.com

[153] https://www.google.com/adsense

[154] Churchill, C. (2004)

[155] http://www.kanoodle.com

[156] Thurow, S. (2004)

[157] Sullivan, D. (2000)

[158] Zerdick, A. et al. (2001), p. 27

[159] Oser, K. (2002)

[160] Sullivan, D. (2003a)

Major search engines can also gain revenues from *licensing*. Google, for example, has gained substantial revenue from licensing technology to portals and directories like AOL and Yahoo.[161]

The current indexes of major search engines include more than four billion Web pages. The availability of this infrastructure to metasearch engines, niche engines, and innovative agents will surely lead to an expanding market of services for locating information.

[161] Machill, M. et al. (2003), p.20

# 4. The Information Market

Previous chapters provided an overview of technical aspects of search engine architecture. In the search for the optimal design, it is important to take a closer look at circumstances particular to the traded good (information), the media market, and the resulting goals of different stakeholders.

## 4.1. The Information Economy

Content is the main resource of the Information Society. Through query ranking, search engines determine which documents will be displayed. This places them in a position of power and social responsibility,[162] an Internet bottleneck reducing the flood of information to a manageable volume.[163]

In the past, publishing information was expensive, a financial risk limiting production.[164] The ease and affordability of Internet publishing has led to a diminished overall quality.[165] Content providers are challenged to design new business models linking editorial products and corporate communication without threatening credibility.[166]

As internal agents helping users sort through the wealth of material available online,[167] search engines must confront two problems: *selection* and *interpretation*. Selection can be managed with ranking algorithms as discussed previously. The need for interpretation occurs with increasing amounts of information and a declining fraction of meaningful content.

The Internet triggered an explosion of information. As an *experience good*,[168] its relevance cannot be judged prior to consumption. Even then, judging its value can be difficult. Search engines, like other media producers, seek to build reputation and establish brand loyalty as a substitute for knowledge about quality.

The search engine user behaves as a 'Satisficer' or 'rational ignorasmus', accepting offers in the form of top search results and disregarding the rest.

## 4.2. Transaction Costs

In contrast to the neoclassical economic model, characterized by complete transparency with perfect and free information, the market for information is

---

[162] Introna, L., Nissenbaum, H. (2000), pp. 169-171

[163] Machill, M. et al. (2002), p. 8

[164] Kiefer, M. (2001), Medienökonomik, pp. 352-356

[165] Compared to traditional mass media, Web publishing entails minimal costs. Readers can break out of their passive role and become journalists themselves.

[166] Machill, M. et al. (2002), p. 17

[167] In contrast, a printed TV guide is an example of an *intermedia* product that helps readers cope with the information flood of another medium, television.

[168] Shapiro, C., Varian, H. (1999), p. 5

associated with *transaction costs* that raise the price of products. Transaction costs can be divided into search, decision, and control costs.

The demand for information depends in part on its benefits, while the propensity to extend a search increases with:

- transaction value (relative to a total budget)
- actual or assumed price dispersion
- higher user aversion to risk
- lower individual search costs
- lower urgency of demand
- lower-rated experiences with alternative offers.[169]

A recipient cannot check all potentially relevant material and wants reliable sources. Transaction costs can be lowered by sorting a collection of information, providing it with a clear structure, and enhancing its credibility.

Professional editors (like those at Yahoo) can reduce the volume and increase the quality of a collection of documents (sorting). Kleinberg's HITS algorithm is designed to find authority and hub sites relevant to a query (credibility). Google[170] uses the DMOZ directory[171] to establish a clear structure and sort sites by importance (structure).[172]

### 4.3. Economics of Attention

In the traditional economic model, the production and distribution of goods is determined by the allocation of limited resources. Our society has moved to an economy based largely on information, a resource typically not subject to scarcity. In fact, we are in many ways facing a flood of information that is often difficult to manage. A key limited resource of the Information Economy is *attention*; the 'New Economy' could be described as the 'Attention Economy'.[173] User attention is the most important Internet resource.[174] Information without attention has no impact. Media complement, compete, and replace each other but the user's attention stays limited.

Companies selling products or services online depend on the attention[175] they receive from visitors.[176] Their Web sites should offer free content and be well pre-

---

[169] Zerdick, A. et al. (2001), pp. 39-42

[170] http://www.google.com

[171] http://www.dmoz.org

[172] This is measured with PageRank.

[173] Zerdick, A. et al. (2001), pp. 36-37

[174] Weichert, S. (2003), p. 1

[175] Attention can be measured by *stickiness* and expressed in one of three ways: the total time a visitor spends on a site, the number of visits per user; or the number of pages viewed.

[176] Davenport, T., Beck, J. (2001), pp. 114-116

pared and easily found. Search engines address the last claim in particular, allowing users to search the Web and select relevant documents.[177] Search engines direct the attention of users, taking on the function of a powerful gatekeeper.[178]

### 4.4. Mass Media

Like traditional mass media, search engines provide information; finding the right information is the primary concern of search engine users.[179] But, in contrast to traditional media, search engines do not produce content, but rather provide information about the location of documents containing desired content, directing user attention accordingly.

Traditional media industries are characterized by relatively high *market concentration* resulting from economies of scale and scope.[180] *Economies of scale* are the major incentive to horizontal concentration in the media sector. These economies are characterized by non-rivalness of consumption. The average cost of production falls as output rises. The cost of the first copy of a newspaper, for instance, is very high because of high fixed costs. Costs per unit fall with a rising print run. In the case of digital goods, this phenomenon is even more pronounced because nearly all production costs accrue to the first copy, while distribution costs are minimal. The cost of generating an active index of more than four billion Web pages is very high, while the cost for this infrastructure is fixed. This creates advantages for large companies and leads to concentration in the media market.

*Economies of scope* are the major incentive to vertical and diagonal concentration among media outlets. *Vertical concentration* occurs because non-rival resources (e.g., expertise and R & D activities) can be applied across market segments within a medium to achieve an improved cost structure. *Diagonal concentration* is made desirable by exploiting this opportunity in more than one medium.[181]

Journalistic production and advertisement is characterized by economies of scope in *production*, *distribution*, and *consumption*. For example, the printing, delivery, and attention-drawing costs of a newspaper's journalistic production can be used to subsidize advertising.[182]

---

[177] Meckel, M. (2003), pp. 8-9

[178] Machill, M. et al. (2003), p. 18

[179] Kiefer, M. (2001), pp. 53-57

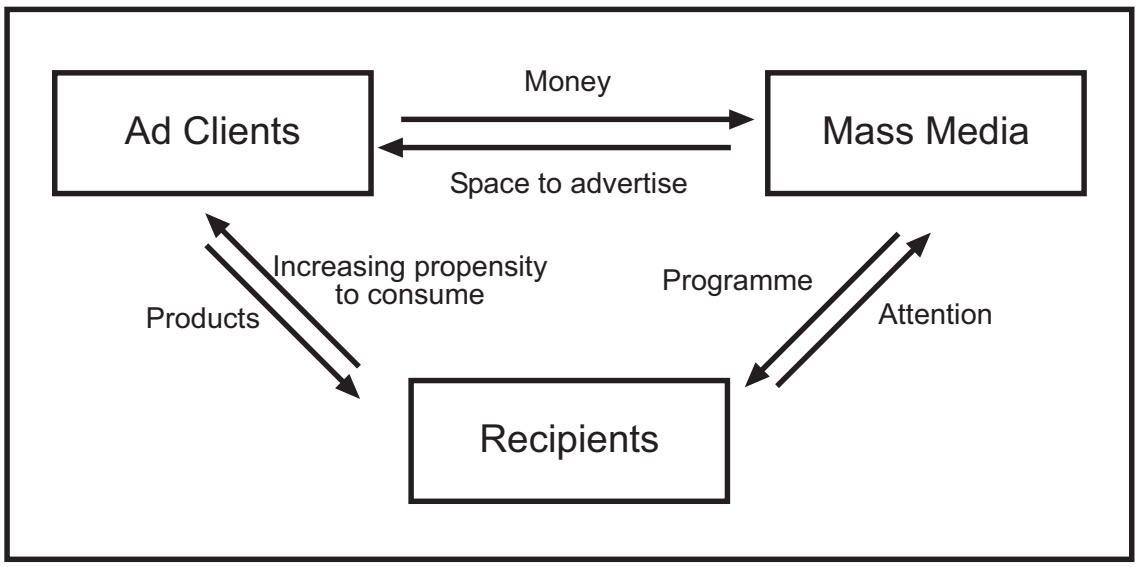
[180] Kops, M. (1999), pp. 1-4

[181] A newspaper, for example, could extend its activities by acquiring a competitor (horizontal concentration), founding its own distribution company (vertical concentration), or launching an Internet portal (diagonal concentration).

[182] In *production*, for instance, there are costs for the printing machine that is needed to print the editorial part. The additional printing of advertisement does not cause high additional costs. The *distribution* network of a newspaper can also be used for advertisement with low extra costs. The attention that is created by the products of the journalist can also be used for advertisement in the *consumption*.

The following figure shows the connection between producers, ad clients, and recipients in traditional advertising-financed mass media associated with economies of scale and scope.

Figure 7:
Trilateral Relationship in Traditional Advertising-Financed Media



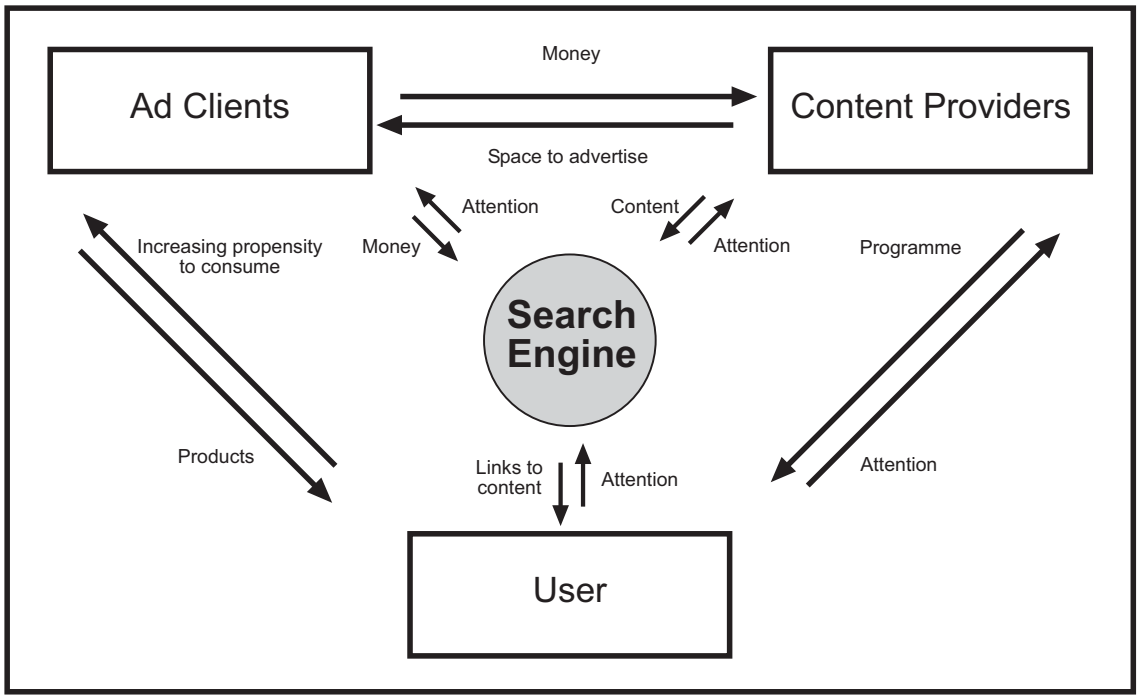

Source: based on Zerdick, A. et al. (2001), p. 50

This relation indicates that in selling the gained attention of recipients, advertising-financed mass media will always consider the interests of ad clients and not simply work to satisfy recipient needs for entertainment, relaxation, information, etc. For the recipient, the programme seems to be free, but the actual circumstances are more complex.[183]

This same pattern is found on the Internet. Search engines help users manage the large volume of information and retrieve documents sorted by relevance to a query. The majority of users are not willing to pay for content or search services. Major search engines benefit from economies of scope and gain revenues by selling paid links from ad clients. The following figure illustrates how an advertising-financed search engine directs the attention of users to content providers and ad clients.

[183] Zerdick, A. et al. (2001), p. 50

Figure 8:
Trilateral Relationship of Advertising-Financed Search Engines




Source: own representation

### 4.5. Diversity of Opinions

Search engines select and judge the relevancy of information, providing them an important and powerful role in society. As with traditional media, one can legitimately ask if the information goods of new media should somehow be regulated[184] to protect the public interest.[185]

Since ad clients pay for search services, providers adjust to their requirements, creating a potential for market failure as performance from the perspective of the user deteriorates.[186] Users must always keep in mind that major search engines are privately held companies with no obligation to serve the public interest.[187]

Using a major search engine, it is very likely that users will be directed to large sites whose designers have the technical knowledge to win the ranking game. Users are less likely to find small and less popular sites not optimised by professionals. In the current commercial model, search engines seeking to realize greatest popularity would tend to provide for the majority of interests. For a

[184] In Germany, the content of private broadcasting is required to represent a diversity of opinions. Minority views are addressed and important political, ideological, and social groups should have an opportunity to express themselves so that public opinion is not influenced in an unbalanced way. See Kops, M. (1999), p. 11

[185] Kops, M. (2000), pp. 35-39

[186] Machill, M. et al. (2003), pp. 441-442

[187] Hargittai, E. (2004)

search engine, the cost of losing a small group of users may be outweighed by the benefits of addressing the masses and those paying for the various forms of enhanced visibility.[188]

Media and other cultural institutions should be aware of this problem and monitor search engines in their role as powerful gatekeepers. Users either are not interested or do not understand the problems that arise from concentration in the search engine market. Therefore, media should work to improve user awareness.[189]

In contrast to traditional media, search engines do not make decisions about the production of content.[190] New ranking techniques, like the personalized or social network ranking discussed previously, rank Web sites by user interest and offer hope of solving this problem by turning the representation of a diversity of opinions into a competitive advantage.

[188] Introna, L., Nissenbaum, H. (2000), pp. 175-177

[189] Salkever, A. (2003)

[190] Traditional media pay for broadcast programming and it is often not profitable to buy or produce a programme addressing minorities. On the Internet, a wide variety of opinion is represented on Web sites accessed for free. The task of a search engine is to direct the users to sites of interest.

# 5. Search Engines from an Agency Theory Perspective

## 5.1. The Basic Concept of Agency Theory

Together with Transaction Cost and Property Rights Theories,[191] Agency Theory is an element of Organizational Economics, an interdisciplinary perspective developed to explain the structure, power, and efficiency of institutions.[192] It describes a contractual framework in which cooperative effort is often compromised by opportunistic behavior. This problem can often be resolved through incentive systems and control structures.

### 5.1.1. Main Concept and Assumptions

In agency theory, individuals are assumed to be rational, risk-averse, and motivated by self-interest, whereas organizations are characterized by goal incongruence and information asymmetry.[193] Information is seen as a commodity in a relationship in which the income of one party (a principal) depends on the behavior of another (an agent). Principals control the means of production, delegating authority to agents, and can choose to expend resources to obtain better information about agent behavior.[194]

This structure has inherent problems.[195] Agent goals may differ from those of principals (goal incongruence), principals cannot fully observe the actions of agents and may not be aware of information influencing their behavior (information asymmetry), and the parties may have different attitudes toward risk assessment.[196] Agency theory seeks to resolve these conflicts by applying suitable incentive schemes and control mechanisms.[197]

Unfortunately, additional costs develop when a principal attempts to control agent misbehavior.[198] Both principals and agents have an incentive to reduce these costs because benefits from savings can be shared between the two parties. Thus, there is a common interest to define a monitoring and incentive structure that limits costs associated with information exchange.[199]

---

[191] Kiefer, M. (2001), pp. 54-57

[192] Saam, N. J. (2002), p. 5

[193] Karake-Shalhoub, Z. (2002), p. 101

[194] Child, J., Faulkner, D. (2002), p. 23

[195] Clegg, S. et al. (1996), pp. 124-125

[196] Demougin, D., Jost, P. (2001), pp. 23-24

[197] Karake-Shalhoub, Z. (2002), p. 102

[198] Heinrich, J. (2001), pp. 186-187

[199] Clegg, S. et al. (1996), p. 125

### 5.1.2. Agency Problems

Problems that arise in principal-agent relationships can be classified into four categories, summarized in the table below.[200]

Figure 9:
Agency Problems

| | **Hidden Characteristics** | **Hidden Intentions** | **Hidden Knowledge** | **Hidden Action** |
|---|---|---|---|---|
| **Point of time** | Before the contract signing | After the contract signing | After the contract signing | After the contract signing |
| **Behavior (Interpretation)** | Exogenously given (Qualification) | Dependent on willingness (Fairness) | Exogenously given (Knowledge) | Dependent on willingness (Effort) |
| **Transparency of Behavior** | Known ex post | Known ex post | Ex post hidden | Ex post hidden |

Source: based on Saam, N. J. (2002), p. 30

*Hidden characteristics* are present before the signing of a contract because an agent has private information about his own characteristics (e.g., misrepresented qualifications). *Hidden intentions* develop because an agent can act unfairly and 'hold up' a principal who cannot easily withdraw from a contract because of sunk costs. *Hidden knowledge* (also known as hidden information) is similar to hidden characteristics, but relates instead to circumstances that develop during the execution of a contract which can be exploited by an agent. After a contract is completed, an agent can misrepresent his work in ways that cannot easily be observed or judged by a principal. This behavior is called *hidden action*.

### 5.1.3. Solutions to Agency Problems

There are a variety of common solutions for each of these agency problems.[201]

- *Monitoring* and information systems, directed at hidden action problems, discipline agents by keeping principals well-informed about agent activities.
- The importance of *incentives* as a means to address hidden information and hidden action problems increases as monitoring becomes less feasible. An efficient incentive compensation system allows agents to benefit from increased productivity and provides a common objective. However, the implementation of an incentive system itself leads to additional costs, diminishing the expected higher returns. Incentive systems are typically used to solve hidden information and hidden action problems.

[200] Saam, N. J. (2002), pp. 28-31; Picot, A. et al. (2001), pp. 57-61
[201] Ibid., pp. 31-35

- *Vertical integration* between the parties allows principals to direct agents with a contractual relationship. Partnership is replaced by hierarchy with the possibility of sanctions. An employment contract is a typical example of vertical integration that is applied to solve hidden intention problems.
- Principals can induce agents to reveal hidden characteristics by *self-selection*, allowing them to choose from a set of contracts.
- In *signaling*, an agent reveals hidden characteristics to obtain increased compensation.
- Hidden action can be reduced through a *bonding* system under which agents agree to limits on their behavior and sanctions in case of non-fulfillment.
- In *screening*, principals improve their selection process through, for example, instruments such as assessment centers.

### 5.2. The Optimal Delegation of Power

In the Information Society, knowledge expands rapidly. The resulting flood of information leads to a greater emphasis on specialization and a division of labor, with an increasing reliance on the judgment of experts.[202] Principal-agent-relationships result from the utilization of the benefits of division of labor.[203] Delegating power to an agent has both advantages and disadvantages. Judgments about extent of such delegation can be critical to success in business.[204]

Agents have special skills that allow principals to conserve resources (e.g., time). In Figure 10, the marginal benefits of delegation are expressed by the MB curve. The first unit of time delegated to an agent yields greater marginal cost savings than the last. The marginal cost of a delegation of power (the C curve) represents both control costs and external costs and increases with expanded delegation.[205] The optimal level of delegation is found at D*, where a further increase of delegation would lead to more costs than benefits for the principal. At that point, the marginal costs and marginal benefits of delegation are equal.

---

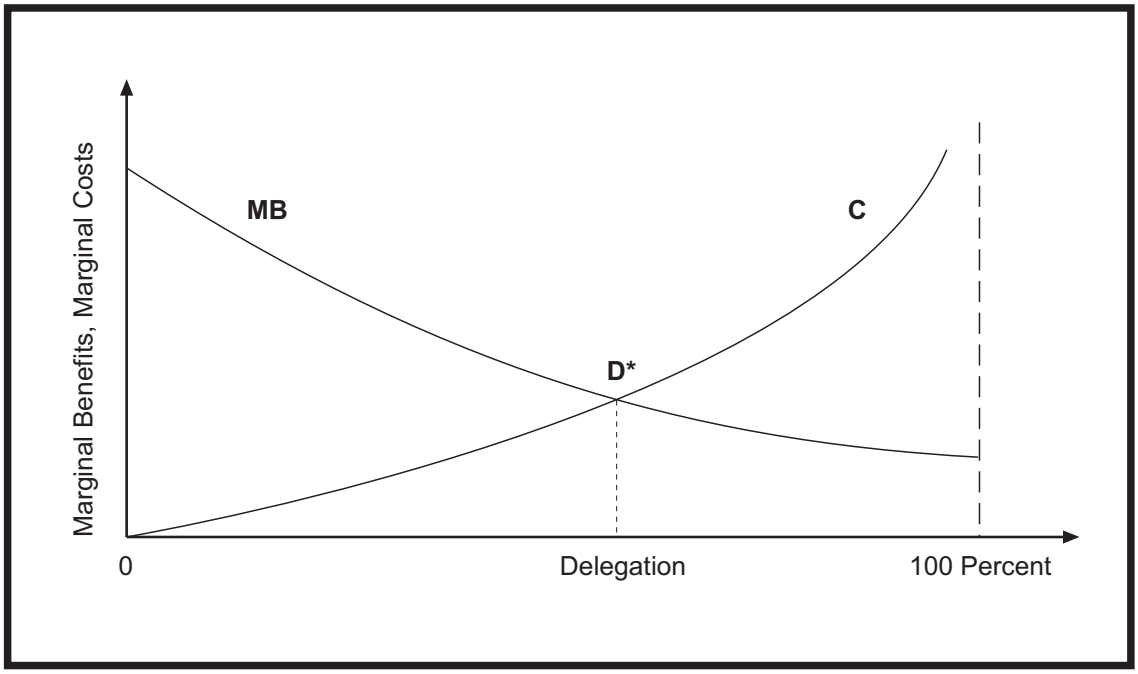
[202] Kops, M. (1998), p. 5

[203] Jost, P. (2001), p. 1

[204] Blankart, C. (1994), pp. 274-279

[205] An agent is able to act more autonomously with a rising degree of delegation. This deviation from the instructions of the principal yields external costs that have to be added to control costs. These external costs rise with declining possibility to judge the quality of the product, benefits from specialization and level of delegation to the agent. See Kops, M. (1996), p. 9

Figure 10:
Optimal Level of Delegation




Source: Blankart, C. (1994), p. 276

### 5.3. Power Roles

Collecting and presenting links to information on the Internet is a process that involves different organizations and industries. Contemplating the optimal design of a search engine requires an understanding of the goals and needs of all those interacting groups. The next section presents a framework for strategic thinking based on Turow's analysis of power roles in mass media industries.[206] The defined power roles illuminate the needs of the various groups, which could be described as goal-directed, boundary-maintaining[207] activity systems.

The groups need resources from the environment.[208] A search engine, for example, has to rely on enterprises like content or network providers for resources. In this context, power involves applying resources to bring others to agreement.[209] Groups themselves are not power roles; they carry out activities related to a power role. Thus, a group can assume more than one role within the infor-

[206] Turow, J. (1992), pp. 19-51

[207] Not everyone is allowed to participate in the group. For example, only editors employed by Yahoo can decide whether a site will be included in the company's directory and how its content will be described.

[208] Resources are the people, supplies, information, services, and money gathered from outside the boundaries of the group.

[209] An example is a strike by a union. Unions rely on their control of labour resources to obtain higher wages. Not all exertions of power are so evident; often, the most effective are those not perceived by the people being influenced.

mation retrieval industry. A company acting as a content provider, for example, can also act as an ad client or a search engine user.

The division of labor between search engines and other groups generates costs and benefits. The search for the optimal search engine design involves finding the simultaneously optimal levels of delegation between search engines and all interacting groups. A variety of power roles can be identified in the search engine industry.

Directories like Yahoo employ *editors* who decide which web sites will be included, classifying them into different categories and writing descriptions of their pages. Listed sites benefit directly from visitors browsing the directory and indirectly from a higher ranking in crawler-based search engines.[210] With search engines as principals delegating classification to editors as agents, problems like hidden action or hidden characteristics can develop. For example, an editor can do sloppy work or unjustly promote sites for reasons of personal interest. It is important to find the optimal level of delegation between directories (principals) and editors (agents).[211] This raises questions about the assignment of tasks to one or more editors and also about their compensation.[212] *Job enlargement*, where a team of editors is responsible for a number of categories, is a possible solution.

*Authority* represents a society as a whole and can influence search engines with general conditions. Regulations and self-imposed obligations should increase the welfare of a society.[213] Racist, violent, pornographic, or anti-democratic content can have negative external effects on the society.[214] In 2003, search engines appeared for the first time in the German law that requires a representative for the protection of children and young people.[215] The establishment of a voluntary code of conduct is a signal from search engine operators to both legislators and users. It demonstrates that they are accepting social responsibility and conserving resources that otherwise would have been expended to prescribe and enforce regulations. A more reliable signal from a code of conduct leads to lower control costs, a higher optimal level of delegation, and greater benefits resulting from a division of labor. Procedures to make the use of search engines more transparent are an important element of self-regulation. Users want to know whether a link is sponsored, what filtering takes place, and

---

[210] Web sites listed in important directories are considered to be more relevant by many ranking techniques. Furthermore, important directories like Yahoo or DMOZ are often the initial point from which crawlers start indexing. See Thelwall, M. (2001), p. 117

[211] For example, the harder it is to monitor an editor, the higher the control costs will be for a search engine, limiting optimal delegation.

[212] Kräkel, M., Sliwka, D. (2001), pp. 331-333

[213] In contrast, China's decision to deny access to Google may well be diminishing that nation's social welfare. The decision was based on censorship considerations; Google's cache feature allows users to view blocked sites. This clearly illustrates the potential power of authority.

[214] Kops, M. (2000), p. 22

[215] Welp, C. (2003), p. 491

how results are ranked.[216] Selfregulation[217] is a research project at Oxford University investigating current self-regulatory codes of conduct.

*Investors* control the financial resources of a search engine, seeking to maximize its market value. Examples for the influence of investors are the concern that Google could sacrifice long-term opportunities to meet quarterly market expectations because of outside pressure from the shareholders after the IPO[218] and the rumor about changes in Google's ranking algorithm for the 'Florida' update before the company's forthcoming IPO.[219] Search engine optimizers speculated that, in order to win investors, Google might have set up a filter to generate irregular rankings and create a big turnover in the AdWords program. Imperfect markets with asymmetric information between capital seekers (search engines) and capital providers (investors) create risks resulting from the potential behavior of capital seekers. Uncertainties about quality and behavior can lead to agency problems. For example, investors can be cheated systematically by capital seekers who have more information than they do about the expected revenues of a search engine.[220] This hidden characteristics problem can be reduced by contractual incentives or by monitoring. To limit monitoring costs, this task can be delegated to an institution that acts on behalf of all investors. A bank is an example of such a financial intermediary. Investors delegate the monitoring of the search engine to the bank and share the monitoring costs.[221] The optimal level of delegation increases as control costs decline.

*Licensees* purchase search engine technology. Google's search results, for example, are shown on portals like AOL and Netscape.[222] Niche and meta-search engines can use the infrastructure (index) of a major search engine to display or enrich results according to user needs. In this scenario, the licensee is a principal employing the technology of a major search engine (agent). Information asymmetries exist with regard to both the quality of the index and the ranking techniques. The resulting control costs can be reduced by third-party trust makers that have special knowledge to judge quality (screening). Marginal benefits will equal marginal control costs at the optimal level of delegation. The search engine can invest in a reputable brand name as a signal for quality. This can lower control costs, increase the optimal level of delegation, and maximize the benefits from a division of labor.

The relationship among the three stakeholders (users, ad clients, and content providers) is determined by the willingness of users to pay, potential advertising revenues, and the goals and capabilities of the search engine.

---

[216] Alexander, M. (2003)

[217] http://www.selfregulation.info

[218] Schmidt, E. (2004), p. i

[219] Gupta, A. (2003)

[220] Hartmann-Wendels, T. (2001), pp. 117-121

[221] Ibid (2001), pp. 144-145

[222] Sullivan, D. (2004c)

The next section outlines two design patterns representing possible relationships among stakeholders.

## 5.4. Design Patterns

Design patterns are the basis for strategies that determine a search engine's field of activity, business model, and relationship to stakeholders. These patterns follow from criteria discussed previously.

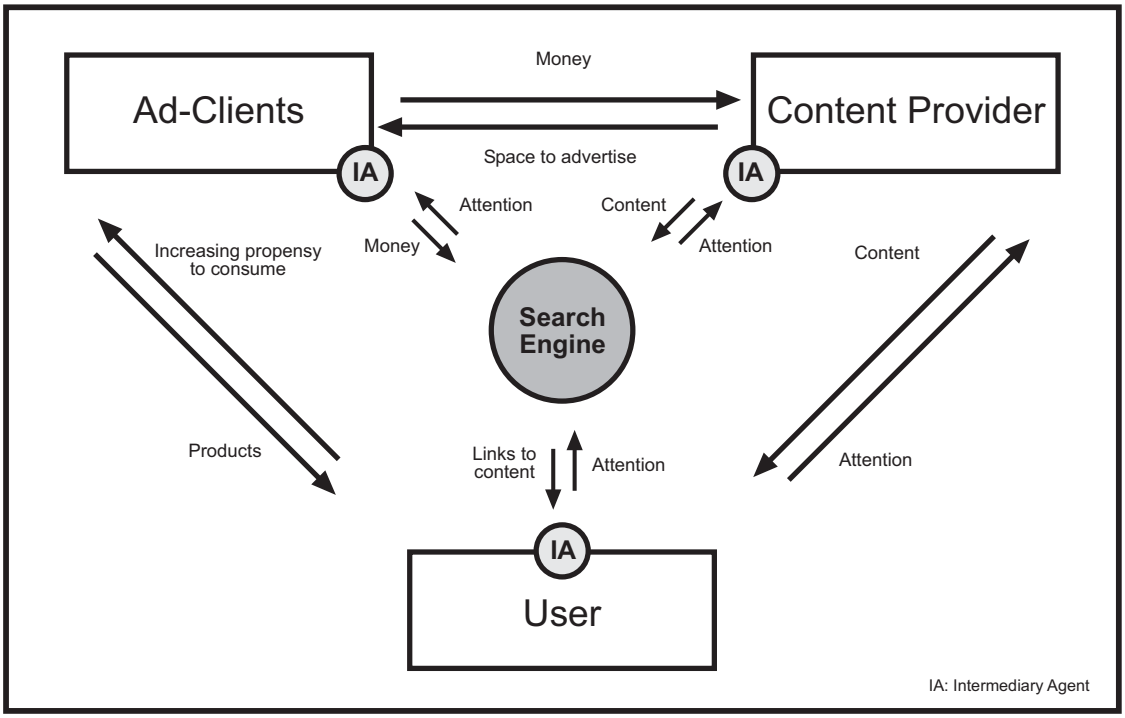
- *technical aspects* of search engine architecture, ranking techniques, and trends in Web search
- the *nature of information*, the information market, and media industry
- theoretical tenets of *Agency Theory* that elucidate stakeholder relationships.

Different design patterns suggest different strategies. In contrast to a trilateral model, a direct relationship is not influenced by the interests of third parties like ad clients.

### 5.4.1. A Trilateral Relationship Pattern

Most search engines are financed by advertising, operating within a structure represented by the pattern illustrated in the following figure.

Figure 11:
Trilateral Relationship




Source: own representation

*Content providers* need to gain attention for their products. Their market place challenge is not information access, but rather information overload.[223] The critical resource on the Internet is attention. It is easy to set up a Web site; the difficulty lies in developing viewership. Search engines help users find valuable information, directing their attention to content providers who want a transparent ranking of the search result because this knowledge enables them to optimize their documents. Since most users do not want to pay for content, providers try to earn money with advertising, sales, or through business models such as content syndication.

*Ad clients* want to maximize the attention from attracted visitors within a given budget. They seek visitors with a strong interest in the content offered and want a central hub where all costs can be controlled. Search engines have the same concern: paid listings should be closely related to a query. In their desire to maximize revenue per keyword,[224] they want clients to write relevant descriptions of their Web sites.

Search engine *users* generally are not aware of ranking techniques. In fact, ranking algorithms are in most cases protected secrets. Users focus instead on receiving valid results that rank documents by relevance, without concern for link sponsorship, retrieval techniques, or business models.[225] Users also demand that results be clearly arranged and quickly retrieved, preferably with an intuitive query that does not require powerful but unfamiliar features or search languages. They tend to submit simple queries with one or two words and hope to find something relevant. Search engines that focus more on algorithms than on usability fail to meet the needs of the users.[226] Users want a ranking that is not influenced by ad clients or other stakeholders. Unlabeled paid listings and spam should not be included in editorial results and a user-friendly interface should be provided.

*Search engines*, on the other hand, aim to generate revenue, concealing algorithms to protect their economic value and avoid spamming. They want to be seen as independent and competent gatekeeper, directing user attention to high-quality content. Most users have a preferred search engine. (Google is currently the most popular.) Surveys indicate that about half employ at least one alternative, likely a niche engine for special demands. This suggests that developers might profit from targeting niches that are at present not adequately served.[227]

Because of information asymmetry and resulting agency costs, the use of specialized *intermediary agents* is a common strategy. Typical intermediary agents employed by *users* include metasearch engines that combine results from different search engines with a uniform search language, social search engines that

---

[223] Shapiro, C., Varian, H. (1999), pp. 6-7

[224] Search engines hold an auction for keywords, ranking sponsored links accordingly.

[225] Machill, M. et al. (2002), pp. 7-8

[226] Machill, M. et al. (2003), p. 440

[227] Ibid (2003), pp. 443-444

use a community to personalize results, graphical user interfaces that group results according to semantics and trust makers that apply special knowledge to screen results.

Knowledge Agents are intermediary agents that are situated between the user and the search engine and specialize in a specific domain by extracting characteristic information from search results. Queries are refined according to the domain specific knowledge to find most relevant documents for a given query within a domain of interest.[228] Query Tracker[229] is an intermediary agent that uses Google's index. Users submit queries that are run through the search engine every 24 hours. Through feedback, the system creates a user profile that improves the relevance of returns and generates personalized queries automatically. Poorly formed queries are expanded and filtered according to the profile.[230]

*Intermediary agents* used by *content providers* include optimisers that influence site design to allow for easy indexing of documents that will then rank high in the result for given keywords.

Pay per performance search engines are an example of *intermediary agents* used by *ad clients* to purchase paid links.

Agency Problems and Possible Solutions

In a search for information, the wealth of material available on the Internet today often makes it impossible to visit all potentially relevant Web sites. A division of labour to a search engine (agent) can significantly lower search costs for users (principals).

The goals of *user and search engine* are congruent in many respects. Both search engines and users seek to avoid spam, achieve fast retrieval, and operate within an intuitive interface design. In addition to these congruent goals, however, there are information asymmetries and incongruent goals that result in agency costs. For example, search engines have more knowledge about ranking algorithms. Understanding these disparities is a critical factor in identifying agency problems. The main goal of a commercial search engine is profit. It gains revenue by directing user attention to content providers that pay for traffic sent to a site. In contrast, users want to find documents most relevant to a query. This can lead to a hidden intention problem for users, one that can be reduced by monitoring, signaling, and greater transparency in ranking.

For users, efficiency is gained by delegating *monitoring* to trusted third parties (e.g., professional journals) that have special knowledge. Another approach is the *signaling*, wherein a search engine invests in a brand name. The goal is to make reputation a signal for quality. In the long run, this investment will be unprofitable (and generate sunk costs) if advertised features are not perceived as authentic. Finally, transparency can be enhanced by the publication of ranking

[228] Aridor, Y. et al. (2000), pp. 15-16

[229] http://www.cs.colostate.edu/~somlo/QueryTracker/

[230] Anthers, G. H. (2004)

algorithms and labelling of sponsored links. Certificates, ratings, and tests can complement active screening by users.[231] The cost of signaling will typically rise as the quality of search results declines.

With increasing delegation of power from the user, a search engine has more opportunity to decide autonomously and control costs increase. The optimal level of delegation is the point at which marginal costs equal marginal benefits.[232] Marginal benefits from delegation will increase with the relevance of retrieved documents, whereas marginal costs are determined by the degree of transparency.

Search engines earn revenues by indexing high-quality content and directing users to ad clients. Users will visit a search engine repeatedly only if they are satisfied with the results. Content providers want their Web sites indexed by search engines to gain the attention of directed visitors.

*Search engines* (principals) invite *content providers* (agents) to make documents available for crawling. A content provider can exclude his site from indexing or, conversely, work to facilitate the process. The exclusion of all or part of a site can easily be accomplished with a file called 'robot.txt' placed in the root of the domain[233]; crawling can be expedited by means of optimization. A content provider will logically optimize if the resulting costs are less than the benefits accrued through increased viewership.

Content providers can try to spam the index of search engines to gain attention after submission and indexing and cause hidden intention problems. These problems can be resolved with the application of monitoring and bonding systems. Most search engines detect (monitoring) and punish (bonding) spammers by banning their sites from being indexed for a period of time. Many publish guidelines for webmasters that describe how Web sites should be designed for indexing and what is considered spam (transparency).

Monitoring, bonding, and increasing transparency create agency costs. The optimal level of a delegation of authority to content providers is the point at which these marginal costs equal the associated marginal benefits. Marginal benefits of delegation rise for Web sites containing high-quality content prepared for crawling, whereas marginal costs rise with increased spamming.

*Ad clients* (principals) want *search engines* (agents) to direct user attention to their Web site. As agents, search engines have more information about directed visitors and can act opportunistically against the interests of ad clients. Problems of asymmetric information can occur before and after the signing of a contract. Ad clients can attempt to control search engines to some degree by analyzing log files, but this entails monitoring costs and, moreover, provides no guarantee of not being cheated. Trusted monitoring system and bonding sys-

[231] Eggs, H. (2001), pp. 93-94

[232] Blankart, C. (1994), pp. 273-277

[233] This file tells the crawler which documents it may spider (download).

tems, where an agent's fee is contingent upon performance, are potential solutions to this problem.

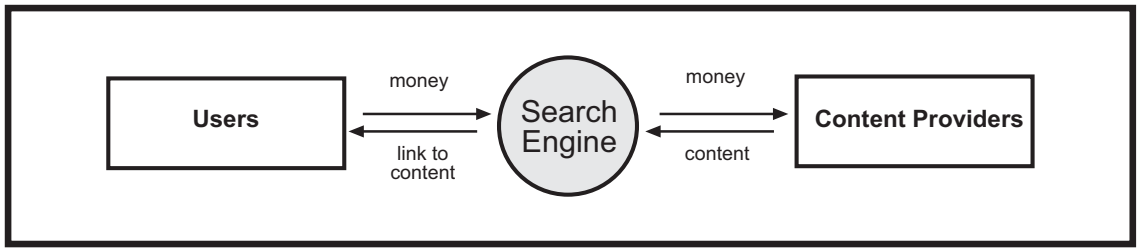
Marginal benefits from a division of labor can be increased with tools that support ad clients in the production of keyword ads. Many paid-placement search engines, for example, provide a keyword suggestion tool. Marginal costs of a division of labor fall with greater transparency. The optimal level of delegation is the point at which marginal costs equal marginal benefits.

In the trilateral model, search engines are agents to both users and ad clients. At the same time, to satisfy user needs, they act as a principal to content providers. To maximize the benefits from a division of labor, search engines must find the optimal levels of delegation with the other stakeholders simultaneously.

### 5.4.2. Direct Relationship Pattern

Direct relationship strategies are designed for small groups of users willing to pay for a service not biased by the interests of a third party. Commercial search engines adopt this pattern only if the inclusion of advertising disturbs users to such a degree that reduced revenues from a willingness to pay cannot be compensated for by the potential advertising revenues. Search engines financed by research programs are more likely to choose this design pattern as illustrated by the following figure.

Figure 12:
Direct Relationship




Source: own representation

From an agency theory perspective, the advantage of this relationship is that the search engine is an agent for the user only. Thus, the goals of search engines and users are not as disparate as they are in the trilateral model. To maximize the benefits of a division of labor, search engines can reduce agency costs by employing the same instruments applicable to the trilateral model. In the relationship with users, these are monitoring (trust maker), signaling (branding) and greater transparency (ranking-rules). In the relation with content providers, these are monitoring (spam detection), bonding (index exclusion), and greater transparency (ranking-rules and webmaster-guidelines). The optimal level of delegation is the point at which marginal costs equal marginal benefits.

## 5.5. Design Strategies

Individuals differ in their activities, interests, opinions, and demographic characteristics.[234] This leads to different expectations, preferences, and patterns of use regarding search engines. The patterns outlined above provide a framework for discussing strategies of optimal search engine design. They determine the choice of business model and field of activity, and the can be implemented individually or in combination.

### 5.5.1. Concentration Strategy

Special circumstances distinguishing information retrieval in particular and media markets in general underlie the substantial amount of concentration found in the search engine industry. Conventional business strategies leading to concentration reflect the trilateral design pattern discussed earlier. Economies of scale make it profitable to maximize the number of users accessing an index.

Only a few search engines can afford to produce an index. Most are set on top of the infrastructure of the major operators. Ask,[235] for example, is based on the Teoma[236] index. The field of activity is extended to the intermediaries of users and ad clients, whereas the importance of intermediate agents to content providers can be diminished by increasing transparency.[237]

This concentration strategy can be elucidated by examining recent actions of Google and Yahoo, two important search engines. Figure 13 illustrates Google's design from an agency theory perspective.

Google used market power to expand its agency activities from the user side to the 'connected' stakeholders—ad clients and content providers. This made it possible to substitute existing agents and acquire new means of earning revenues.

---

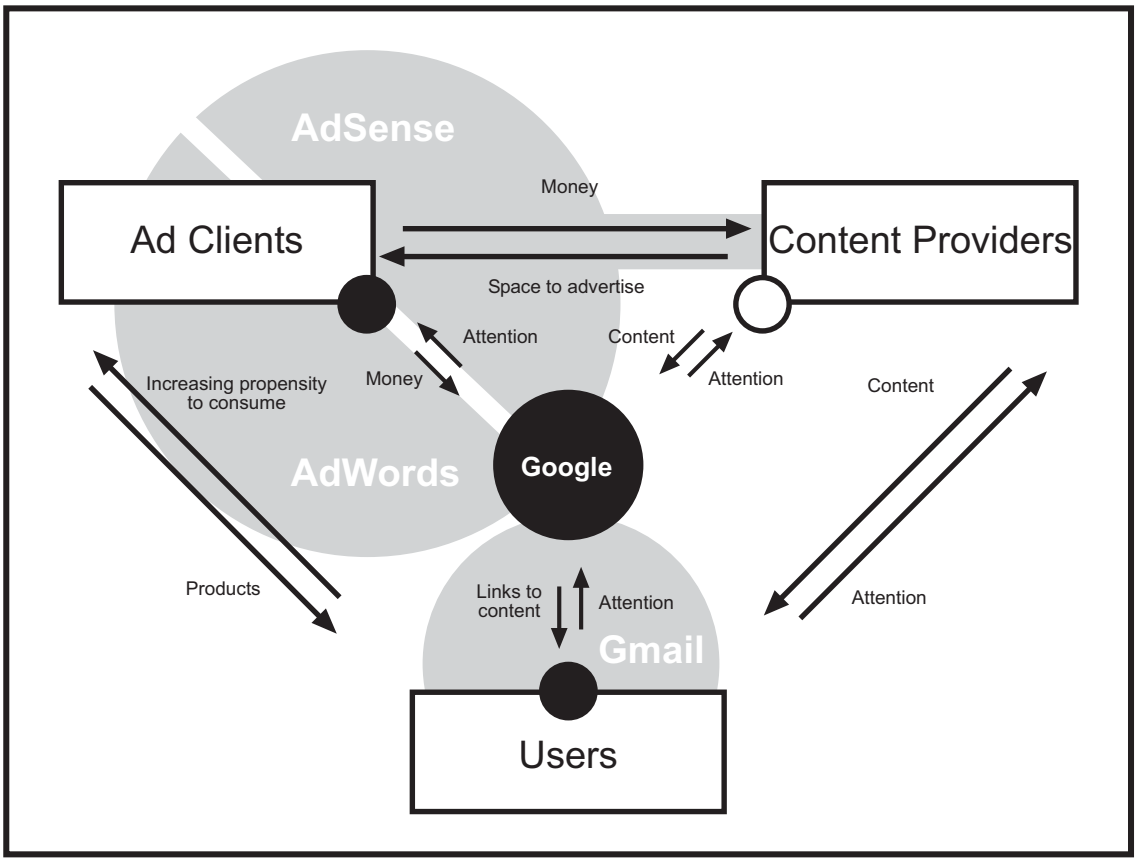
[234] Koppelmann, U. (2000), p. 75

[235] http://ask.com/

[236] http://www.teoma.com/

[237] Search engine optimizers are an example of intermediate agents to content providers. Their market power is weakened by search engines that publish ranking rules and guidelines for designing Web sites that can easily be crawled.

Figure 13:
Google and a Concentration Strategy




Source: own representation

A) Google between Users and Content Providers

The traditional role of a search engine is to act as an agent on behalf of users who want relevant information in response to a query. To meet this demand, Google provides an interface that supports over 97 languages and is easy to use for the majority of users. Some of the popular portals, by adding ‘sticky’ features designed to keep visitors on-site, fail to accomplish this.

The DMOZ directory offers an enhanced presentation that lists the sites in a category by rank rather than alphabetical order. Experienced users can employ the advanced search interface that enables them to pose complex queries using tools such as Boolean operators, topic relevant filtering, thesauri, query reformulation, and query extension. They can search for definitions, file types, and similar pages. Google has integrated specialized databases, such as phonebooks, stock quote listings, and street maps, and included others related to topics ranging from travel to patents, establishing a foundation that will be important for future developments like local search. It is easy to customize a search with parameters like preferred language, number of retrieved documents, or adult filter. It is very likely that the number of these features will increase with new retrieval techniques. Google Labs tests new developments, like a wireless shopping search engine and voice search. Gmail, the company’s new freemail,

is their first 'sticky' feature, keeping users on-site and allowing for the collection of information about users that can be valuable in tailoring personalized search.

Google has anticipated the advantages of economies of scale and scope by investing in the industry's largest index, achieving preeminence in market share. Through Google API, over four billion Web pages can be accessed by developers and researchers building applications. Froogle, for example, is a shopping comparison search engine operated by Google that uses the existing infrastructure and brand to enter new markets.

B) Google between Users and Ad Clients

Paid-placement search engines, e.g., Overture and Espotting, are intermediate agents that simplify the inclusion of paid links for ad clients in various listings. Google's substantial market share and established brand allowed the company to launch its own program for the inclusion of paid listings. AdWords is a substitute for third-party paid-placement search engines that has developed into an important source of revenue. Clients can target keyword ads to specific countries and languages. Budget forecasts and a set of online tools allow them to monitor the inclusion of ads.

A recent change in their ranking of paid listings diminishes incongruent goals from an agency theory perspective. Ad clients want to maximize visitor attention with a given budget, whereas Google wants to maximize its revenues (price per click multiplied by the number of clicks) with relevant ads that do not annoy users. Before the change, the ranking of paid listings was determined by the price per click, regardless of how well an ad matched the keyword. The ranking is now determined by a combination of the price per click and the number of clicks. Ads less relevant to users will be clicked less often, causing them to move down the list. This creates an incentive for clients to produce ads that match well with keywords. Relevant ads are less offensive to users and more effective at inducing them to choose paid listings over editorial ones. In this way, both search engine and ad client goals are achieved. Additionally, users benefit from ads with greater relevance.

C) Google between Ad Clients and Content Providers

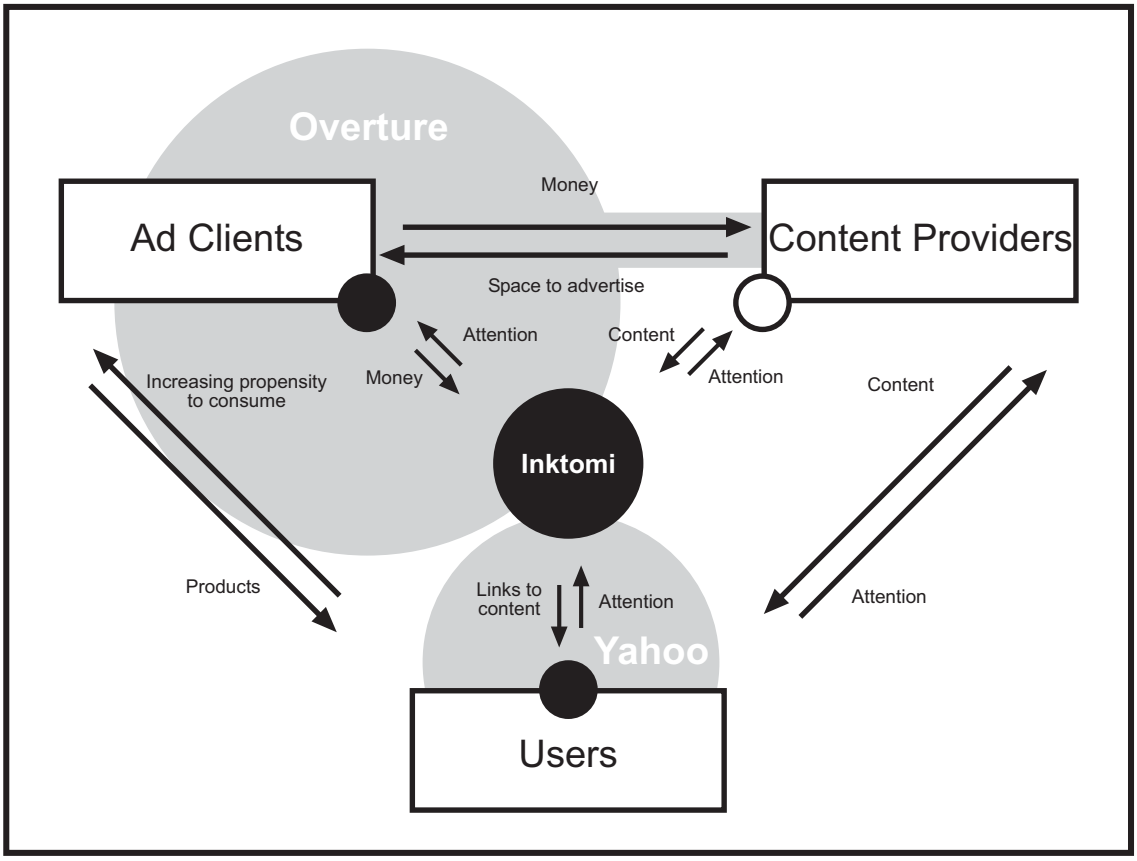
The success of the AdWords programme led Google to extend that concept. AdSense includes paid listings from the AdWord programme in the sites of content providers.

Google publishes information that increases its transparency. This includes guidelines for webmasters, offering advice on how to design sites to allow for easy indexing and alerting them to those techniques regarded as spam. A Google representative (known as 'GoogleGuy') is available on a popular discussion board[238] to decrease information asymmetries regarding optimization.

[238] http://www.webmasterworld.com/

Yahoo is another search engine that has adopted a strategy designed to benefit from market concentration. Recent changes in Yahoo's architecture (illustrated in Figure 11 can be interpreted as an implementation of concentration strategy.

Figure 14:
Yahoo and a Concentration Strategy




Source: own representation

Yahoo is a directory that in the past included Google's search results. Paid listings from Overture provided a counterpart to Google's AdWords and AdSense. In July 2003, Yahoo bought Overture and, after also acquiring Inktomi (a major search provider), replaced Google's results with its own in February 2004.[239] By maintaining a presentation very similar to the familiar Google results, the site has retained a stable viewership.[240]

These acquisitions have made Yahoo an important independent player in the information retrieval industry, a market that has developed into a highly concentrated oligopoly. Users tend to employ only one or two search engines, a behavior the major service providers encourage by supplying search utilities (toolbars) for browsers.[241] Nearly 50 % of respondents to a recent survey have in-

[239] Sherman, C. (2004)
[240] Sullivan, D. (2004d)
[241] Alexander, M. (2003)

stalled at least one toolbar: 22 % use Yahoo's, followed by Google's 20 %, and MSN's 17 %.[242]

### 5.5.2. Niche Strategy

Like a concentration strategy, a niche strategy also follows from the trilateral design pattern. In this case, providers address themselves to a smaller number of users without expanding their field of activity. Instead, they operate as intermediate agents based on the infrastructure of a major provider, working to meet specialized needs.

These can be classified into several categories:

- There is surprisingly little overlap[243] among major service providers, even for popular search terms.[244] Metasearch engines (e.g., Dogpile[245]) conduct a search on different major indexes, a process referred to as *index expansion*.
- *Deep Web access* allows users to search documents hidden in specialized databases. ProFusion[246] is an example of this type of service provider.
- *Vertical search engines* (also known as niche search engines) focus on a limited set of topics.
- *News search engines* allow users to retrieve news stories from hundreds of sources across the Web. All major search engines have their own news-related utility, but there are many other engines that focus on specialized niches, such as newspapers, news feeds, blogs, magazines, and regional news.[247]
- *Shopping comparison* search engines allow users to locate shops by category and check prices at various online stores.
- *Multimedia* search engines retrieve sound, image, and video files, as well as radio and television programs. SpeechBot,[248] for example, has indexed over 17,000 hours of content that can be searched with speech recognition software.

Niche search engines are financed with advertising and are therefore subject to the same problems of incongruent goals and information asymmetries that characterize all business models based on the trilateral relationship pattern.[249]

---

[242] Sherman, C. (2004a)

[243] Overlap refers to a document being listed in more than one index. This concept can be visualized with the Thumbshots Ranking Tool (http://ranking.thumbshots.com/).

[244] Sherman, C. (2004c)

[245] http://www.dogpile.com/

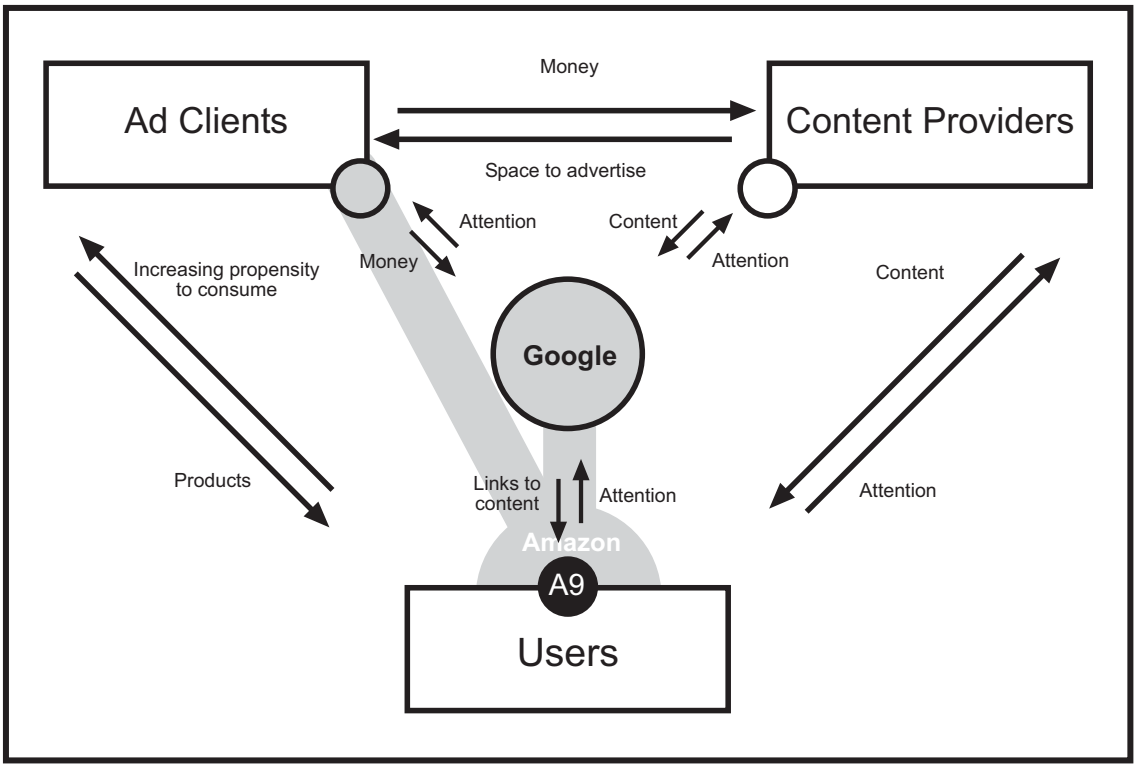
[246] http://www.profusion.com/

[247] Sullivan, D. (2003c)

[248] http://speechbot.research.compaq.com/

[249] See 0for a description of agency problems and solutions to them.

Figure 15:
A9 and a Niche Strategy




Source: own representation

Amazon's A9[250] is a niche search engine based on Google's index that addresses to the group of Amazon customers[251] and includes a number of unique additional features allowing users to personalize results based on their search history and other information available from Amazon.[252] 'Search inside the book' allows users to view scanned pages from more than 120,000 books included with results. 'People who visit this page also visit' is another potentially valuable feature. There is also an option to install a toolbar that provides a pop-up blocker, a dictionary, a thesaurus, and a diary for recording notes about sites that will reappear during subsequent visits. The collection of personal data raises privacy concerns; users must decide if they trust Amazon and desire to benefit from its improved ranking and additional information or if they view the collection of personal information as troublesome. A9 provides a privacy policy clearly stating what personal information is collected and how it is stored. This is a signal to reduce information asymmetries and privacy concerns. Personalization can improve the relevancy of search results; trusted agents with existing user profiles like Amazon are able benefit from it.

[250] http://a9.com/

[251] Users log on the search engine with their Amazon usernames. Information about a user's interests that is recorded when browsing Amazon's site can also be used to further personalize search results.

[252] Sherman, C. (2004d)

### 5.5.3 Bundling Strategy

Bundling combines two or more products or services into a package. Articles are bundled as journals, shampoo can be bundled with conditioner, and a computer is a bundle of software and hardware. Through bundling, firms can increase revenues, reduce costs, and raise barriers to market entry. In the case of digital goods (e.g., digital music or software), bundling is highly effective because these items cost very little to reproduce. Bundling has powerful implications for the structure of those markets in which large firms have a substantial competitive advantage.[253] It is the business logic behind Microsoft's decision to sell Word, Excel, and PowerPoint as an 'Office' bundle.

Another form of bundling, one that is widely considered violative of anti-trust legislation, characterized that company's approach to sales of Windows Media Player. Microsoft used its dominant position in the PC industry to gain market share for its digital media player,[254] the same strategy it employed in establishing Internet Explorer against Netscape in the browser wars.

Microsoft's current search engine (MSN) provides users with paid listings from Overture and editorial listings from LookSmart and Inktomi.[255] It is developing a new search engine that will be integrated with the next version of Windows, using Windows dominant position to bring the search engine into close competition with the current market leaders, Google and Yahoo.[256] This use of a concentration strategy to replace the index and place listings with its own solutions would eliminate the influence of a direct competitor since, as noted, Inktomi and Overture are owned by Yahoo.

Furthermore, Microsoft could also make use of the dominant position of MS Office to reduce the value of the search results of the competitors by limiting the access to proprietary document formats such as Microsoft Word. Microsoft could give its search technology a better ability to search documents of its own file formats or gain revenues from royalties in exchange for giving other search engines the ability to search these documents.[257]

Internet access providers could also make use of a strong position on the market of cable and DSL service providers to enter the search market.[258] These providers have a direct relationship to the users and can easily enrich existing billing-information to user profiles that can additionally be useful for personalized search and tailored ads.

Major service providers may choose to use their strong positions in the search engine market to enter other industries. Hotbot[259] for example, recently introdu-

---

[253] Zhu, K., MacQuarrie, B. (2003), pp. 264-266

[254] Peterson, K. (2004)

[255] Sullivan, D. (2003d)

[256] Borland, J. (2004)

[257] Schmidt, E. (2004), p. 9

[258] Schmidt, E. (2004), p. 4

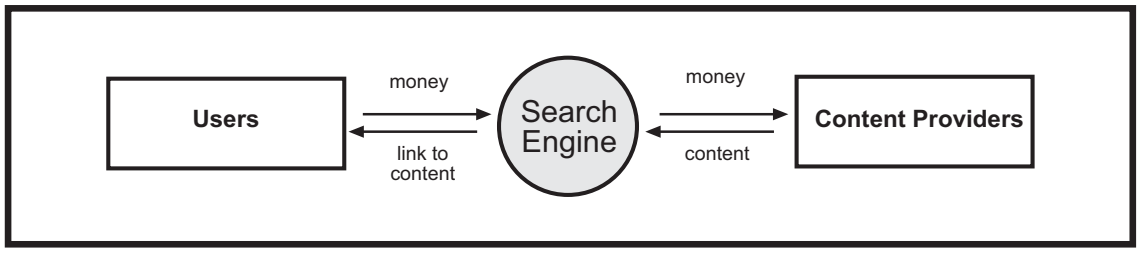
[259] Hotbot (http://www.hotbot.com) is a Lycos-owned search engine that allows access

ced a desktop search utility that allows users to not only search the Web, but also index files and email on a computer, making them searchable as well.[260] Google may launch a similar desktop utility that integrates its toolbar and new email service (Gmail) with storage space.

### 5.5.4. Premium Search

A *premium search* strategy is based on the direct relationship pattern. It addresses a small group of users willing to pay for an unbiased service. As Figure illustrates, the field of activity is centered around user needs.

Figure 16:
Premium Search Strategy




Source: own representation

A commercial search engine will choose this strategy only if the inclusion of advertising lowers the users' willingness to pay to such a degree that adequate revenues cannot be generated. Otherwise, the optimal strategy would likely be a mixture with a niche strategy.[261]

Most public broadcasting Web sites are free of advertising, financed solely through fees paid by households.[262] Because these sites do not serve private commercial interests or produce a return for shareholders, they can focus on providing high-quality content and services, much of which would not otherwise be widely available.

A public broadcasting Internet search service could adopt the direct relationship pattern and compete with commercial providers. A decision to produce such a service would depend on an interpretation of the mandate for public broadcasting.[263] Its extent is largely determined by the resulting costs and its benefit to society.[264] It would need to be controlled by citizens to ensure that it is not abused by government.[265]

---

to the three major indexes (Google, Yahoo (Inktomi), and Teoma). In contrast to metasearch engines, HotBot does not blend results, instead allowing users to switch between the indexes.

[260] Sherman, C. (2004a)

[261] Bhargava, H., Feng, J. (2002), p. 122

[262] Loebbecke et al. (2003), pp. 2-4

[263] Rebmann, R. (2003), pp. 2-4

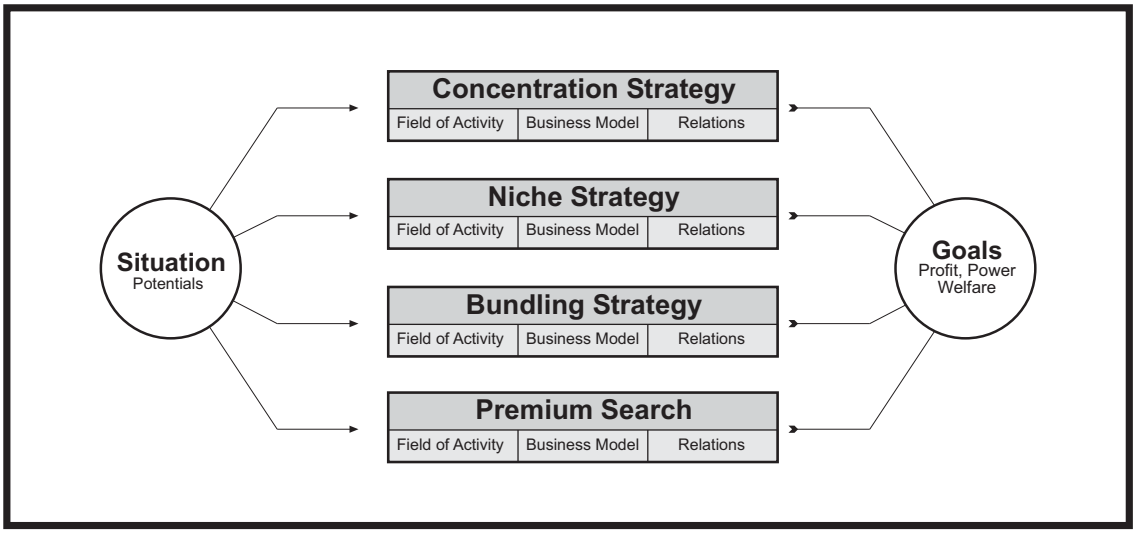
[264] Kops, M. (1998), p. 29

[265] Kops, M. (2000a), p. 1

## 5.6. Choice of Strategy

A search engine service provider's strategic thinking with regard to a business model is influenced by both individual entrepreneurial goals and marketplace limitations. The goals determine a set of strategic choices that in turn is limited by external realities. Strategies can be applied narrowly or in combinations.

Figure 17:
Choice of Strategy




Source: own representation

The development or implementation of a new search technique can dramatically alter a company's strategic approach. Of course, any fundamental shift in market orientation should be based on a careful analysis of consumer demand for a service and its associated cost structure and revenue potential. Specific details related to relevant market segments and user preferences can be critical to effective strategic planning in this context.

### 5.6.1. Goals and Compatible Strategies

The design of a search engine is determined mainly by provider goals. The usual goal is profit, but other possible goals[266], like power or welfare, can influence the strategic choices that define the field of activity, the business model, and relationships to stakeholders.

[266] Koppelmann, U. (2000), pp. 251-253

Figure 18:
Goals and Strategies

| | Profit | Influence | Common Welfare |
|---|---|---|---|
| Concentration | ● | ● | ● |
| Niche | ● | | ● |
| Premium Search | ● | | ● |
| Bundling | ● | ● | |

Source: own representation

*Profit* is the most prevalent motivation in business. While profit maximization is typically sought, a company may decide to accept lower profits in the short term as a means to expand its market share or revenue base. For instance, a search engine with some customers who pay for the service could decide to offer free service to all customers to attract more users. Campaigns to expand a customer base in this way must be planned and carried out very carefully, however, to make certain a positive outcome is achieved in the end.

All the strategies discussed above are consistent with a goal of profitability. The range of potentially effective strategies is limited by the market potential of a particular search engine design and the needs of users. For example, a start-up company that has invented a new retrieval technique will not have the financial resources to choose a concentration strategy. User preferences will determine if it is possible to charge for the service (premium search strategy) or if the service has to be financed with revenues from advertising (niche strategy).

Strategies chosen because of an *influence goal* attempt to strengthen the assets and potential of a service provider. A company could decide to decline a risky investment that has substantial profit opportunity because of safety concerns or to ward off a hostile takeover. For example, Yahoo's concentration strategy (the acquisition of Inktomi and Overture) might not maximize profits, especially in the short term, but does ensure the company will remain an important player in the search engine market.

Another important factor in choosing influence as a goal is market power. In establishing a strong position for Windows Media Player and Internet Explorer, Microsoft used its market power to set product standards and suppress competition in new markets through bundling.

Economic outcomes that benefit society as a whole achieve *common welfare goals*. A government agency or other public sector institution could launch a search engine to compete with major commercial engines but focus on an unbiased retrieval of documents, allowing all views to be represented. The infrastructure could be made freely available for innovation and the development of

new techniques. To avoid having the service influenced by ad clients, the financing could be funded through tax revenues or some other form of public sponsorship.[267] Besides offering a diversity of opinions, public sector sponsorship of a search service could be used to diminish information asymmetries, with projects like a job search engine.

Common welfare goals can also be achieved with an integration of ad clients in either a concentration or niche strategy. The success of Linux demonstrates that the open source[268] idea is viable. Any manipulation in favor of a third party is unlikely because the ranking techniques are controlled by a community of developers. Published ranking rules lower information asymmetries. Advertising can be labeled as such and the revenues can be used to finance an indexing infrastructure able to compete with major search engines. That same infrastructure can then be made available to start-ups and research facilities to develop new retrieval techniques that expand and improve Web search technology. Nutch[269] is an example of an open source search engine project (not yet open to the public) working in a demonstration mode with 100 million pages.[270] Mozdex[271] is another open source engine based on Nutch and seeded with URLs from DMOZ. Its 'explain' feature, which provides ranking details, holds considerable promise.

### 5.6.2. Circumstances Restricting Choice

Not all strategies can be matched to goals and executed successfully. The choice is effectively limited by the search engine's potential as determined by characteristics of the audience of users. Figure 16 provides an overview of different design strategies and the circumstances required for their implementation. The first three potentialities – size, research, and market power – are determined by the service provider. The last – willingness to pay – is determined by the target group the search engine addresses. A strategy or a combination of strategies can be implemented.

---

[267] A service requiring direct payments from users means that a minority would be paying for a premium, unbiased search engine. If the goal is to increase the welfare of the whole society, the service must be free in order to attract as many users as possible.

[268] The code underlying open source software is freely available to anyone and can be extended or modified. Innovations are contributed to the free code base.

[269] http://www.nutch.org/

[270] Battelle, J. (2003)

[271] http://www.mozdex.com/

Figure 19:
Circumstances Restricting Choice

| | Size / Fin. Resources | Innovative Research | Market Power | Willingness to Pay for Service |
|---|---|---|---|---|
| **Concentration** | ● | | | |
| **Niche** | | ● | | |
| **Premium Search** | | ● | | ● |
| **Bundling** | ● | | ● | |

Source: own representation

As in traditional media, the powerful leverage provided by economies of scale and scope creates a strong tendency toward *concentration* in the search engine market. A substantial amount of resources is required to build and maintain an up-to-date index and develop the software needed to search it. There are only a few search engines that can afford their own index. Thus, market share and financial resources are very important factors for achieving success in a concentration strategy. The majority of users do not want to pay for search services. A search engine that wants to attract a lot of users must offer its service free of charge and earn revenues from paid listings. A willingness to pay is not required because collecting user fees would simply diminish the potential customer base. The concentration strategy is viable only for major search engines with the required resources. A strong position within the search engine market can be used to expand and take over the business of intermediate agents.

Small and innovative search engines can follow the *niche strategy*, building on the index of a major service provider. This strategy demands an innovative retrieval technique that offers enhanced benefits to a small user group. A willingness to pay, among users who demand returns not influenced by third parties, distinguishes the niche strategy from *premium search*.

For a commercial search engine, revenues determine the choice of strategy. If there is an adequate willingness to pay despite the inclusion of advertising, a service provider can choose a mixture of both strategies. The degree of inclusion determines the balance between revenues drawn from advertising and those accumulated through user fees, with the optimal level maximizing revenues.

The *bundling* strategy can be implemented effectively only by companies with both sufficient financial resources and a dominant position in another market. Microsoft, for example, has the resources to acquire or build its own infrastructure and integrate a search engine into Windows. This could allow Microsoft to use its dominant position in the operating systems market to gain a strong position in the search market.

# 6. Conclusions

This paper attempts to answer questions about the optimal design of search engines by providing an overview of technical aspects of search engine architecture and trends in their design (chapters 2 and 3), outlining circumstances that affect the information market and media industries (chapter 4), and exploring potential benefits that result from a delegation of labor through insights gained from agency theory (chapter 5). Based on that discussion, four design strategies that determine a search engine's business model, field of activity, and relationships to stakeholders have been examined.

A *concentration strategy* addresses a large number of users and earns revenues from advertising. Search engines that follow this strategy have their own index and in many cases contract for its use with smaller service providers that cannot afford to produce one. Companies with sufficient financial resources expand their field of activity to paid-placement search engines. A *niche strategy* addresses a smaller number of users that have specialized needs. The service is often based on the index of a major provider, activities are highly focused on user needs, and operations are financed through advertising. *Premium search* also serves a small market but is financed by user fees. Some providers mix this with other strategies. A *bundling strategy* requires a dominant position in another market that can be leveraged in the search engine market.

A service provider's choice of strategy depends on both its goals and external market conditions. Goals establish a set of preferred strategies, while externalities determine the potential viability of strategic choices.

The question of the optimal design of a search engine can be productively posed before the launch of a novel search engine, with the development of a new and promising retrieval technique, or after a successful launch for the purpose of strategic reorientation. It is important that market conditions are thoroughly analyzed and that the targeted group of users is well understood. Further suggested research might include the entire product marketing process to find a solution for the systematic development of search engines from an agency theory perspective.

As in traditional media, there are strong concentration tendencies in the search engine market. Yahoo! acquired Inktomi and Overture, in effect 'declaring war' on Google with the replacement of its search results. By means of its forthcoming IPO, Google will acquire substantial financial resources to prepare a counterstrike. With Gmail, Google has introduced its first 'sticky' feature, one that could become quite valuable in generating profiles for personalized search. In addition to Google and Yahoo!, there is a third major player that will inevitably play an important role in the future of search engine services. Microsoft has announced plans to develop its own search technology. As with Windows Media Player and Internet Explorer in the past, that company's overwhelming financial resources and dominant position in the PC market place it in a very strong position to establish a heavyweight search engine service.

Next to the battlefield shared by these three major competitors, there is a space for companies that stand out through innovative search techniques or the integration of unique data. Personalization is a promising trend for these smaller market participants. Cooperative efforts with trusted companies that possess existing user profiles (e.g., Amazon and eBay) have clear advantages when seeking to overcome the privacy concerns of potential users.

The developing competition among the major players and the potential for innovation associated with new retrieval techniques can alter the landscape of the current search engine market considerably. It is not clear who the winners and losers in this game will be. One possible outcome is a breakdown similar to today's television market, with a highly concentrated oligopoly of major players and a group of small special-interest providers.

# Bibliography


AFP: Studie: 2005 surft jeder Zweite mit Breitband-Tempo, http://www.billiger-surfen.de/nachrichten/technik/5139.php3.

Arbeitsgruppe Runder Tisch - Multimedia Home Plattform: Multimedia-Home Plattform (MHP) - Grundlage für die Konvergenz der Medien: Basispapier zum Einstieg in den freien Markt für digitales Fernsehen in Deutschland, hrsg. von der Deutschen TV-Plattform, Frankfurt am Main 1999.

ARD-Forschungsdienst: Digitales und interaktives Fernsehen: Nutzererwartungen und Akzeptanzchancen, in: Media Perspektiven, Ausgabe Nr. 8 1999, S. 430-440.

Agosti, M., Melucci, M. (2000), Information Retrieval on the Web, in Lecture Notes in Computer Science, Vol. 1980, Springer Verlag, Berlin Heidelberg, pp. 242- 285

Alexander, M. (2003), PCMLP Self-Regulation Review, http://www.selfregulation.info/iapcoda/0308xx-selfregulation-review.htm (2004-04-19)

Anthers, G. H. (2004), Search Engines get Cleverer, Computer Weekly, http://www.computerweekly.com/articles/article.asp?liArticleID=129800 (2004-04-08)

Arasu, A. et al. (2001), Searching the Web, in: ACM Transactions on Internet Technology, Vol.1/1, pp. 2-43

Aridor, Y. et al. (2000), Knowledge Agents on the Web, in: Lecture Notes in Artificial Intelligence, Vol 1860, Klutsch, M., Kerschberg (Eds.), Springer Verlag, Berlin Heidelberg, pp. 15-26

Babiak, U., (1999), Effekive Suche im Internet, O'Reilly Verlag, Köln

Baeza-Yates, R., Ribeiro-Netro, B. (1999), Modern Information Retrieval, Addison-Wesley, ACM Press, New York

Baeza-Yates, R. et al. (2002), Web Structure, Dynamics and Page Quality, in: Lecture Notes in Computer Science (2476), Laender, A., Oliveira, A. (Eds.), Springer Verlag, Berlin Heidelberg, pp. 117-130

Baeza-Yates, R. (2003), Information Retrieval in the Web Beyond Current Search Engines, in: International Journal of Approximate Reasoning, Vol. 34, pp. 97-104

Battelle, J. (2003), An Open Source Search Engine, http://searchenginewatch.com/searchday/article.php/3071971 (2004-04-16)

Bergstein, B. (2004), Net Searches Getting Sharper Focus, Contra Costa Times, http://www.contracostatimes.com/mld/cctimes/business/7684762.htm (2004-02-02)

Berners-Lee, T. et al. (2001), The Semantic Web, http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21 (2004-01-03)

Bhargava, H., Feng, J. (2002), Paid Placement Strategies for Internet Search Engines, in: Proceedings of the eleventh international conference on World Wide Web, pp. 117-123

Blankart, C. (1994), Club Governments versus Representative Governments, in: Constitutional Political Economy, Vol. 3/1994, pp. 273-286

Borland, J. (2004), Microsoft's Long-Playing Business Record, http://zdnet.com.com/2100-1104-5190698.html (2004-04-24)

Bozsak, E. et al. (2002), KAON – Towards a Large Scale Semantic Web, in: Lecture Notes in Computer Science (2455), Bauknecht, K. et al. (Eds.), Springer Verlag, Berlin Heidelberg, pp. 304-312

Bradman, O. et al. (2000), Crawler-Friendly Web Servers, in: ACM SIGMETRICS Performance Evaluation Review, Vol. 28/2, pp. 9-14

Brin, S., Page, L. (1998), The Anatomy of a Large-Scale Hypertextual Web Search Engine, http://dbpubs.stanford.edu:8090/pub/1998-8, (2003-09-22)

Chang, G. et al. (2001), Mining the World Wide Web, Kluwer Academic Publishers, Boston

Chau, M., Chen, H. (2003), Personalized and Focused Web Spiders, in: Web Intelligence, Zhong, N. et al. (Eds.), Springer Verlag, Berlin Heidelberg, pp. 197-216

Chen, Y. et al. (2002), I/O-Efficient Techniques for Computing Pagerank, in: Proceedings of the eleventh international conference on information and knowledge management, pp. 549-557

Child, J., Faulkner, D. (2002), Strategies of Cooperation, Managing Alliances, Networks and Joint Ventures, Oxford University Press, Oxford

Choi, O. et al. (2003), Semantic Web Search Model for Information Retrieval of the Semantic Data, in: Lecture Notes in Computer Science (2713), Chung, C. et al. (Eds.), Springer Verlag, Berlin Heidelberg, pp. 588-593

Chowdhury, G., Chowdhury, S. (2001), Searching CD-ROM and Online Information Sources, Library Association Publishing, London

Churchill, C. (2004), Day of Reckoning in Search Engine Advertising, http://searchenginewatch.com/_subscribers/articles/article.php/3298951 (2004-03-19)

Clay, B. (2004), Search Engine Relationship Chart, http://www.bruceclay.com/searchenginerelationshipchart.htm (2004-05-03)

Clegg S. et al. (1996), Handbook of Organization Studies, Sage Publications, London

Craswell, N. et al. (2001), Effective Site Finding using Link Anchor Information, Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 250-257

Davenport, T., Beck, J. (2001), The Attention Economy, Understanding the New Currency of Business, Harvard Business School Press, Boston

Demougin, D., Jost, P. (2001), Die Prinzipal-Agenten Theorie im Unternehmenskontext, in: Die Prinzipal-Agenten Theorie in der Betriebswirtschaftslehre, Jost, P. (Ed.), Schäffer-Poeschel Verlag, Stuttgart, pp. 11-43

Diligenti, M. et al. (2002), Web Page Scoring Systems for Horizontal and Vertical Search, in: Proceedings of the eleventh international conference on World Wide Web, pp. 508-516

Doan, A. et al. (2002), Learning to Map between Ontologies on the Semantic Web, in: Proceedings of the WWW2002

Dvorak, J. C. (2004), Search Engine Mania, The Mad Rush to Develop New Ways of Finding Info Online, PC Magazine, http://abcnews.go.com/sections/scitech/ZDM/web_search_commentary_pcmag_040405.html (2004-04-08)

Eastman, C., Jansen, B. (2003), Coverage, Relevance and Ranking: The Impact of Query Operators on Web Search Engine Results, in: ACM Transactions on Information Systems, Vol. 21/4, pp. 383-411

Eggs, H. (2001), Vertrauen im Electronic Commerce, in: Markt und Unternehmensentwicklung, Picot, A. et al. (Eds.), Deutscher Universitäts-Verlag, Wiesbaden

Endicott, R. C. et al. (2004), Fact Pack 2004, Second Annual Guide to Advertising and Marketing, Supplement to: Advertising Age, Digital Edition: AdAge.com

Ferber, R. (2003), Information Retrieval, dpunkt.verlag, Heidelberg

Fuhr, N. (2000), Models in Information Retrieval, in: Lecture Notes in Computer Science, Vol. 1980, Agosti, M. et al. (Eds.), Springer Verlag, Berlin Heidelberg, pp. 21-50

Gaither, C. (2004), New Search Engine Taps Into Social Networks, The Boston Globe, http://searchenginewatch.com/_subscribers/articles/article.php/3301471 (2004-03-22)

Giles, C. et al. (1998), CiteSeer: An Automatic Citation Indexing System, ACM Press, New York, pp. 89-98

Giles, C. et al. (2003), eBizSearch: A Niche Search Engine for e-Business, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 413-414

Glöggler, M. (2003), Suchmaschinen im Internet, Springer Verlag, Berlin Heidelberg

Goh, D., Ang, R. (2003), Relevancy Rankings: Pay for Performance Search Engines in the Hot Seat, in: Online Information Review, Vol. 27/2, pp. 87-93

Gupta, A. (2003), Google Floria Algo Update, http://www.seorank.com/google-florida-update.htm (2004-02-01)

Hargittai, E. (2004), BBC News, http://news.bbc.co.uk/1/hi/technology/3601371.stm (2004-04-08)

Hartmann-Wendels, T. (2001), Finanzierung, in: Die Prinzipal-Agenten Theorie in der Betriebswirtschaftslehre, Jost, P. (Ed.), Schäffer-Poeschel Verlag, Stuttgart, pp. 117-146

Hawking, D. et al. (1999), Results and Challenges in Web Search Evaluation, Proceedings of the WWW Conference, pp. 244-252

Heinrich, J. (2001), Medienökonomie, Band 1: Mediensystem, Zeitung, Zeitschrift, Anzeigenblatt, Westdeutscher Verlag, Wiesbaden.

Henzinger, M. (2000) Link Analysis in Web Information Retrieval, in: IEEE Data Engineering Bulletin, Vol. 23/3, pp. 3-8

Henzinger, M. (2000a) Web Information Retrieval - an Algorithmic Perspective, in: Proceedings of the 8th Annual European Symposium on Algorithms, pp. 1-8

Henzinger, M. et al. (2002), Challenges in Web Search Engines, in: SIGIR Forum, Vol. 36/2, pp. 11-22

Ingwersen, P. (2000), Users in Context, in: Lecture Notes in Computer Science, Vol. 1980, Agosti, M. et al. (Eds.), Springer Verlag, Berlin Heidelberg, pp. 157-178

Introna, L., Nissenbaum, H. (2000), Shaping the Web: Why the Politics of Search Engines Matters, in: The Information Society, Vol. 16, pp. 169-185

Jeh, G., Widom, J. (2003), Scaling Personalized Web Search, in: Proceedings of the twelfth international conference on World Wide Web, pp. 271-279

John, R., Mooney, G. (2001), Fuzzy User Modeling for Information Retrieval on the World Wide Web, in: Knowledge and Information Systems, Vol. 3, Springer Verlag, London, pp. 81-95

Jost, P. (2001), Die Prinzipal-Agenten Theorie in der Betriebswirtschaftslehre, Schäffer-Poeschel Verlag, Stuttgart

Karake-Shalhoub Z. (2002), Trust and Loyalty in Electronic Commerce, Quorum Books, London

Katz, B. et al. (2002), Omnibase: Uniform Access to Heterogeneous Data for Question Answering, in: Proceedings of the seventh International Workshop on Applications of Natural Language to Information Systems

Kiefer, M. (2001), Medienökonomik, Oldenbourg Verlag, München Wien

Kleinberg, J. (1999), Authoritative Sources in a Hyperlinked Environment, in: Journal of the ACM, Vol 46/5, pp. 604-632

Kline, V. (2002), Missing links: the quest for better search tools, in: Online Information Review, Vol. 26/4, pp. 252-255

Kobayashi, M., Takeda, K. (2000), Information Retrieval on the Web, in: ACM Computing Surveys, Vol. 32/2, pp. 144-173

Koppelmann, U. (2000), Produktmarketing, Springer Verlag, Berlin Heidelberg

Kops, M. (1996), Rechtfertigen von Nachfragemängeln einer Regulierung von Rundfunkprogrammen?, Arbeitspapiere des Instituts für Rundfunkökonomie an der Universität zu Köln, Vol. 72/1998

Kops, M. (1998), Prinzipien der Gestaltung von Rudfunkordnungen, Arbeitspapiere des Instituts für Rundfunkökonomie an der Universität zu Köln, Vol. 100/1998

Kops, M. (1999), Combating Media Concentration in a Globalising Word Economy, Institute for Broadcasting Economics, University of Cologne, Germany, Working Paper No. 118

Kops, M. (2000), Diversifizierte Verfahren zur Bereitstellung von Informationsgütern, Arbeitspapiere des Instituts für Rundfunkökonomie an der Universität zu Köln, Vol. 123

Kops, M. (2000a), Financing and Sustaining Political Will to Support Public Service Broadcasting, Institute for Broadcasting Economics, University of Cologne, Germany, Working Paper No. 121

Kozawa, M. et al. (2002), Constraint Search for Comparison Multiple-Incentive Merchandises, in: Lecture Notes in Computer Science (2455), Bauknecht, K. et al. (Eds.), Springer Verlag, Berlin Heidelberg, pp. 152-161

Kräkel, M., Sliwka, D. (2001), Innerbetriebliche Aufgabenverteilung und Delegation, in: Die Prinzipal-Agenten Theorie in der Betriebswirtschaftslehre, Jost, P. (Ed.), Schäffer-Poeschel Verlag, Stuttgart, pp. 331-357

Kwok, C. et al. (2001), Scaling Question Answering to the Web, in: ACM Transactions on Information Systems, Vol 19/3, pp. 242-262

Leroy, G. et al. (2003), The Use of Dynamic Contexts to Improve Casual Internet Searching, in: ACM Transactions on Information Systems, Vol. 21/3, pp. 229-253

Lin, J. et al. (2003), The Role of Context in Question Answering Systems, in: Proceedings of the 2003 Conference on Human Factors in Computing Systems

Loebbecke et al. (2003), Betriebswirtschaftliche Betrachtung öffentlich-rechtlicher TV-Online Aktivitäten, Arbeitspapiere des Instituts für Rundfunkölonomie an der Universität zu Köln, Vol. 183, pp. 2-4

Ludwig, M. (2003), Breaking Trough the Invisible Web, in: net connect, Winter 2003, pp. 8-10

Machill, M. et al. (2002), Transparenz im Netz, Funktionen und Defizite von Internet-Suchmaschinen, Verlag Bertelsmann Stiftung, Gütersloh

Machill, M. et al. (2003), Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen, in: Wegweiser im Netz, Machill, M., Welp, C. (Eds.), Verlag Bertelsmann Stiftung, Gütersloh

Markoff, J. (2004), Google Planning to Roll Out E-Mail Service, New York Times, http://www.nytimes.com/2004/03/31/technology/31CND-GOOGLE.html?ex=1081569600&en=642bb461d24a7ec5&ei=5070# (2004-04-08)

McGuigan, G. (2003), Invisible Business Information: the Selection on Invisible Web Sites in Construction Subject Pages for Business, in: Collection Building, Vol. 22/2, pp. 68-74

Meckel, M. (2003), Vorwort, in: Wegweiser im Netz, Machill, M., Welp, C. (Eds.), Verlag Bertelsmann Stiftung, Gütersloh

Moens, M. (2000), Automatic Indexing and Abstracting of Documents, Kluwer Academic Publishers, Boston

Moldovan, D., Surdeanu, M. (2003), On the Role of Information Retrieval and Information Extraction in Question Answering Systems, in: Lecture Notes in Artificial Intelligence (2700), Springer Verlag, Berlin Heidelberg, pp. 129-147

Morrissey, B. (2004), Search Engines Eye Personalized Results, Direct Marketers News, http://www.dmnews.com/cgi-bin/artprevbot.cgi?article_id=26723&dest=article (2004-03-06)

Nekrestyanov, I., Panteleeva, N. (2002), Text Retrieval Systems for the Web, in: Programming and Computer Software, Vol. 28/4, pp. 207-225

Nie, J., Chen, J. (2003), Exploiting the Web as Parallel Corpora for Cross-Language Information Retrieval, in: Web Intelligence, Zhong, N. et al. (Eds.), Springer Verlag, Berlin Heidelberg, pp. 218-239

Oser, K. (2002), Yahoo to Send E-Mails Based on Search Engine Activity, http://www.industryclick.com/magnewsarticle.asp?newsarticleid=330086 (2004-04-06)

Parker, P. (2004), Google Tests Personalized Search, Debuts New Look, http://www.clickz.com/news/article.php/3332621 (2004-04-05)

Paulson, J. (2003), Fast Facts About Froogle, SitePoint, http://www.sitepoint.com/print/1060 (2003-09-20)

Peters, C., Sheridan, P. (2000), Multilingual Information Access, in: Lecture Notes in Computer Science, Vol. 1980, Agosti, M. et al. (Eds.), Springer Verlag, Berlin Heidelberg, pp. 51-80

Peterson, K. (2004), EU reports gives an inside look at RealNetworks, http://seattletimes.nwsource.com/html/businesstechnology/2001910377_realnetworks23.html (2004-04-23)

Picot, A. et al. (2001), Die grenzenlose Unternehmung – Information, Organisation und Management, Gabler Verlag, Wiesbaden

Pretto, L. (2002), A Theoretical Analysis of Google's PageRank, in: Lecture Notes in Computer Science (2476), Leander, A., Oliveira, A. (Eds.), Springer Verlag, Berlin Heidelberg, pp. 131-144

Pujol, J. et al. (2003), A Ranking Algorithm Based on Graph Topology to Generate Reputation or Relevance, in: Web Intelligence, Zhong, N. et al. (Eds.), Springer Verlag, Berlin Heidelberg, pp. 381-394

Rappoport, A. (2000), Search Engines: The Hunt is on, Network Computing, CMP, http://www.networkcomputing.com/shared/printArticle.jhtml?article=/1120/1120f1full.html&pub=nwc (2003-11-22)

Reardon, M. (2004), Start-Up Launches Social Search Engine, CNET News.com, http://zdnet.com.com/2100-1104-5144567.html (2004-02-10)

Rebmann, R. (2003), Online-Dienste als wettbewerbswidrige Angebote des öffentlich-rechtlichen Rundfunks, Arbeitspapiere des Instituts für Rundfunkökonomie an der Universität zu Köln, Vol. 182

Richardson, M., Domingos, P. (2002), The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank, in: Advances in Neural Information Processing Systems Vol. 14, MIT Press, Cambridge, pp. 1441-1448

Richardson, M., Domingos, P. (2003), Combining Link and Content Information in Web Search, Web Dynamics, Springer, New York, to appear, http://www.cs.washington.edu/homes/pedrod/ (2003-10-02)

Robertson, S. (2000), Evaluation in Information Retrieval, in: Lecture Notes in Computer Science, Vol. 1980, Agosti, M. et al. (Eds.), Springer Verlag, Berlin Heidelberg, pp. 81-92

Robertson, S. (2002), Comparing the Performance of Adaptive Filtering and Ranked Output Systems, in: Information Retrieval, Vol. 5, pp. 257-268

Rossi, G. et al. (2001), Designing Personalized Web Applications, in: Proceedings of the Tenth International WWW Conference, Hong Kong, pp. 275-284

Saam, N. J. (2002), Prinzipale, Agenten und Macht, Mohr Siebeck, Tübingen

Sadat, F. et al. (2002), Cross-Language Information Retrieval Using Multiple Resources and Combinations for Query Expansion, in: Lecture Notes in Computer Science (2457), Yakhno, T. (Ed.), Springer Verlag, Berlin Heidelberg, 114-122

Salkever, A. (2003), The Web, According to Google, Buisiness Week online, http://www.businessweek.com:/print/technology/content/jun2003/tc20030610_2810_tc104.htm?tc (2004-04-24)

Salkever, A. (2004), Searching for Trouble?, Buisiness Week online, http://www.businessweek.com/technology/content/jan2004/tc20040122_0347_tc047.htm (2004-01-23)

Schmidt, E. (2004), Form S-1 Registration Statement, http://www.sec.gov/Archives/edgar/data/1288776/000119312504073639/ds1.htm, (2004-04-30)

Schimkat, R. et al. (2002), Living Hypertext – Web Retrieval Techniques, in: Lecture Notes in Computer Science (2346), Unger, H. et al. (Eds.), Springer Verlag, Berlin Heidelberg, pp. 1-14

Schwartz, C. (2001), Sorting Out the Web, Ablex Publishing, London

Shah, U. et al. (2002), Information Retrieval on the Semantic Web, Proceedings of Eleventh International Conference of Knowledge Management, ACM, pp. 461-468

Shapiro, C., Varian, H. (1999), Information Rules: A Strategic Guide to the Network Economy, Harvard Business School Press, Boston

Sherman, C. (2003), Help Test the Wondir Search Engine, http://www.searchenginewatch.com/searchday/article.php/2208541 (2004-02-14)

Sherman, C. (2004), Yahoo! Birth of a New Machine, http://searchenginewatch.com/_subscribers/articles/article.php/3314161 (2004-04-24)

Sherman, C. (2004a), HotBot's New Desktop Search Toolbar, http://searchenginewatch.com/searchday/article.php/3339921 (2004-04-29)

Sherman, C. (2004b), Search Engine Users: Loyal or Blase, http://searchenginewatch.com/searchday/article.php/3342041 (2004-04-21)

Sherman, C. (2004c), Exploring Search Engine Overlap, http://searchenginewatch.com/_subscribers/articles/article.php/3346411 (2004-04-29)

Sherman, C. (2004d), Bleeding the Best of Google and Amazon, http://searchenginewatch.com/searchday/article.php/3342881 (2004-04-29)

Silverstein, C. et al. (1998), Analysis of a Very Large Web Search Engine Query Log, Digital SRC, http://www-cs-students.stanford.edu/~csilvers/ (2003-10-06)

Smith, A. (2003), Think local, search global? Comparing search engines for searching geographically specific information, in: Online Information Review, Vol. 27/2, pp. 102-109

Sterling, G. (2004), Local Search: The Hybrid Future, http://searchenginewatch.com/searchday/article.php/3296721 (2004-03-19)

Stuckenschmidt, H. (2002), Approximate Information Filtering on the Semantic Web, in: Lecture Notes in Artificial Intelligence (2479), Jarke, M. et al. (Eds.), Springer Verlag, Berlin Heidelberg, pp. 114-128

Sullivan, D. (2000), Paid Inclusion at Search Engines Gains Ground, http://www.searchenginewatch.com/sereport/article.php/2163151

Sullvian, D. (2003), Major Search Engines and Directories, http://www.searchenginewatch.com/links/article.php/2156221 (2003-10-06)

Sullivan, D. (2003a), Search Privacy at Google & Other Search Engines, http://www.searchenginewatch.com/sereport/article.php/2189531 (2004-04-06)

Sullivan, D. (2003b), Buying Your Way in: Search Engine Advertising Chart, http://www.searchenginewatch.com/webmaster/article.php/2167941 (2004-04-04)

Sullivan, D. (2003c), News Search Engines, http://searchenginewatch.com/links/article.php/2156261 (2004-04-24)

Sullivan, D. (2003d), How MSN Works, http://searchenginewatch.com/_subscribers/article.php/2148981 (2004-04-24)

Sullivan, D. (2004), Google Launches Gmail, a Free Email Service, http://searchenginewatch.com/_subscribers/articles/article.php/3334251 (2004-04-08)

Sullivan, D. (2004a), Eurekster Launches Personalized Social Search, http://searchenginewatch.com/_subscribers/articles/article.php/3301471 (2004-03-10)

Sullivan, D. (2004b), Google Releases Orkut Social Networking Service, http://searchenginewatch.com/searchday/article.php/3302741 (2004-04-08)

Sullivan, D. (2004c), Who Powers Whom? Search Providers Chart, http://searchenginewatch.com/reports/print.php/34701_2156401 (2004-03-14)

Sullivan, D. (2004d), Google Tops, But MSN Growth Good & Yahoo Switch A Success So Far, http://searchenginewatch.com/_subscribers/articles/article.php/3334871 (2004-04-24)

Thelwall, M. (2001), Commercial Web site links, in: Internet Research: Electronic Networking Applications and Policy, Vol. 11/2, pp. 114-124

Thompson, B. (2003), Is Google too Powerful?, BBC News, http://news.bbc.co.uk/1/hi/technology/2786762.stm (2004-04-01)

Thurow, S. (2004), Kandoodle Joins Contextual Advertising Fray, http://searchenginewatch.com/searchday/article.php/3327651 (2004-03-19)

Turban, E. et al. (2000), Electronic Commerce, Prentice Hall, Upper-Saddle River

Turow, J. (1992), Media Systems in Society, Longman, New York

Walker, Jill (2002), Links and Power: The Political Economy of Linking on the Web, Proceedings of the thirteenth ACM conference on hypertext and hypermedia, pp. 72-73

Weichert, S. (2003), Frisst die Informationsrevolution ihre Kinder?, http://www.politik-digital.de/text/archiv/hintergrund/informationsrevolution.shtml (2004-02-14)

Welp, C. (2003), Ein Code of Conduct für Suchmaschinen, in: Wegweiser im Netz, Machill, M., Welp, C. (Eds.), Verlag Bertelsmann Stiftung, Gütersloh

Wensi, X. et al. (2002), Machine Learning Approach for Homepage Finding Task, in: Lecture Notes in Computer Science (2476), Leander, A., Oliveira, A. (Eds.), Springer Verlag, Berlin Heidelberg, pp. 145-159

Zahdeh, L. (2003), From Search Engines to Question-Answering Systems – The Need For New Tools, http://www.eecs.berkeley.edu/~shawnc/bisctalk_slides/LotfiTalkAbstract.pdf (2003-10-04)

Zerdick, A. et al. (2001), Die Internet Ökonomie, Strategien für die digitale Wirtschaft, Springer Verlag, Berlin Heidelberg

Zhong, N. (2003), Toward Web Intelligence, in: Lecture Notes in Artificial Intelligence, Vol. 2663, Spinger Verlag, Berlin Heidelberg, pp. 1-14

Zhu, K., MacQuarrie, B. (2003), The Economics of Digital Bundling: The Impact of Digitization and Bundling on the Music Industry, in: Communications of the ACM, Vol. 46/9, pp. 264-270

ISSN 0945-8999
ISBN 3-934156-85-1