POLISH-JAPANESE ACADEMY OF INFORMATION TECHNOLOGY

Content Management Systems (CMS)

Lecture 10: Searching for content

Mariusz Trzaska, Ph. D.

mtrzaska@mtrzaska.com, http:// www.mtrzaska.com

Agenda

- Internet challenges
- Technical problems
- Search paradigms
- Information noise
- Smart search engines
- Syntagmatics
- Summary

Materials used from professor K. Subieta

Internet challenges

- Reduce the popularity of traditional newspapers for online newspapers.
- Major changes in phones.
- Major changes in organization and teaching methods.
- Major changes in organization and control of work (enabling home work with full control of the employer).
- Commerce, business, Internet-based administration.

Current status of web resources

- Exponential increase in the number of web pages
- Lower quality of the information presented on them (duplicates, obsolete, banal – lots of garbage!)
- Not always the best way to tailor search methods to user needs and their intellectual capabilities.

Current status of web resources (2)

- "Deep Web" - most of the information available through the Web does not have the form of flat HTML pages, but is stored in databases and is currently mounted as an HTML page in response to user queries.
- This causes the HTML page count to be completely inadequate - there is a virtually infinite combination of information in the databases that can be presented as Web pages.

Current status of web resources (3)

- Although most popular text standards have been absorbed by popular search engines, there are also formats such as audio, graphics, and video that are hardly recognizable and must be explicitly or contextually indexed.
  - This is labor - intensive.
  - Artificial Intelligence methods are not always advanced enough.

Current status of web resources (4)

- More intelligent search methods, based on linguistic engineering, are less effective in terms of the size of Web resources.
- The user is not interested in the information as such, but the information he/she needs to broaden his/her knowledge or make a decision.
  - **Relevance**: The information corresponds formally to a user query.
  - **Pertinency**: information corresponds to the need of the user.

Qualitative changes in Web organization

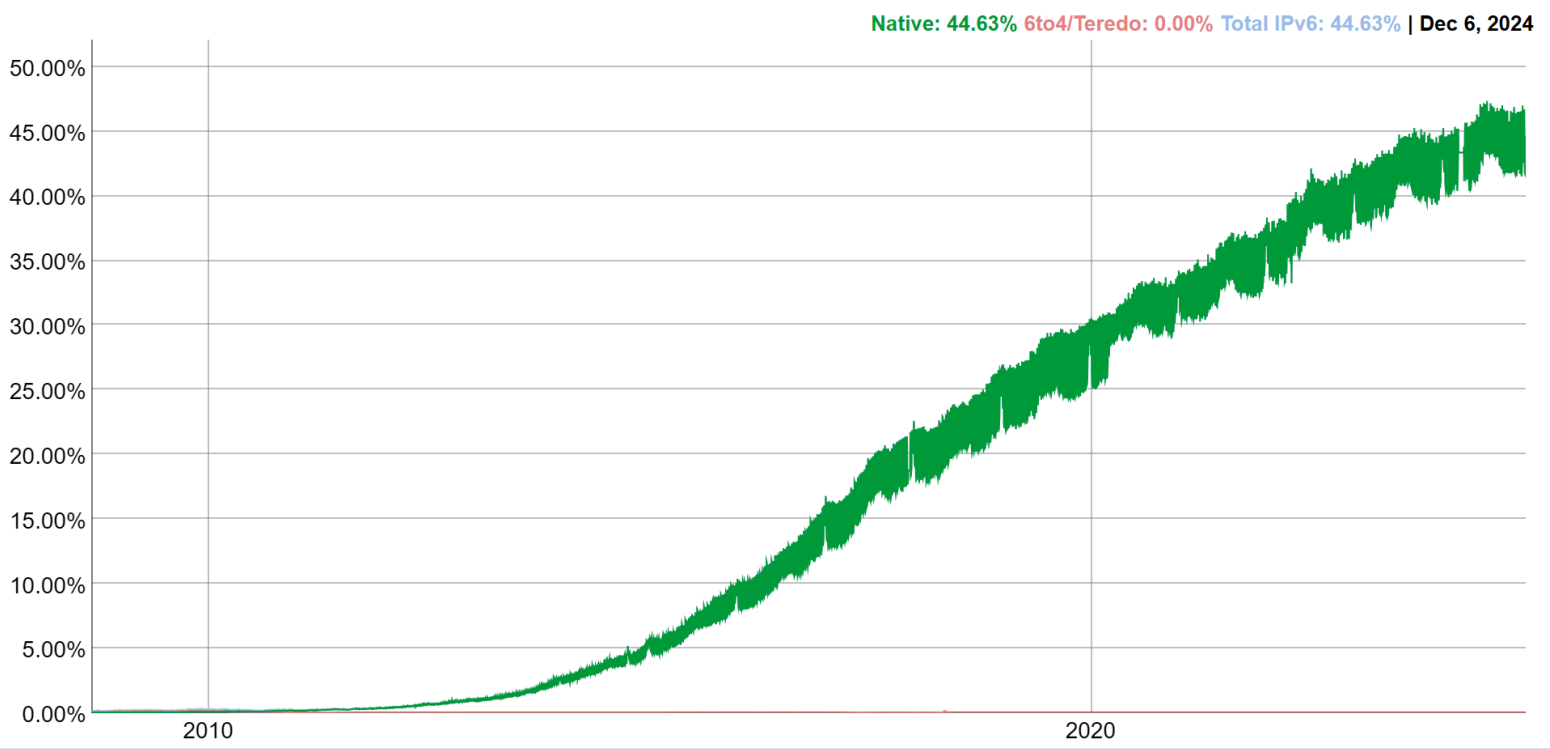
- There are works on the so-called ***semantic web***, which will be Web based on a well-organized database.
- XML is proposed as a structuring tool and a huge R & D flow goes in that direction.
- XML is good as the basis for the standardization of heterogeneous information exchange protocols but is very limited as a data model.

New IP addresses

- Traditional ones are 4 bytes and their stock is running out. They are capable of addressing 256^4, or about 4 billion servers, but additional restrictions make it a much smaller number.
- New IPv6 addresses are 16 bytes, meaning potential for addressing every square centimeter of the earth, for example:
  `2001:0db8:85a3:0042:1000:8a2e:0370:7334`

New IP addresses (2)

- The new protocol is more secure.
- Works with the old version without collision.



Źródło: https://www.google.com/intl/en/ipv6/statistics.html

Website search engines

- The huge size of Web resources makes it necessary to use search engines.
- There is a small number of search engines (Google, Bing, DuckDuckGo, ...) that have proven themselves to be worth and have loyal customers. The rest of the search engines have lost the battle for the market.
- Search engines basically search for keywords in the full text of documents in Web resources.
- So far, there has been no hope of putting a significant "intelligence" into the search engines. These are simple mechanisms.

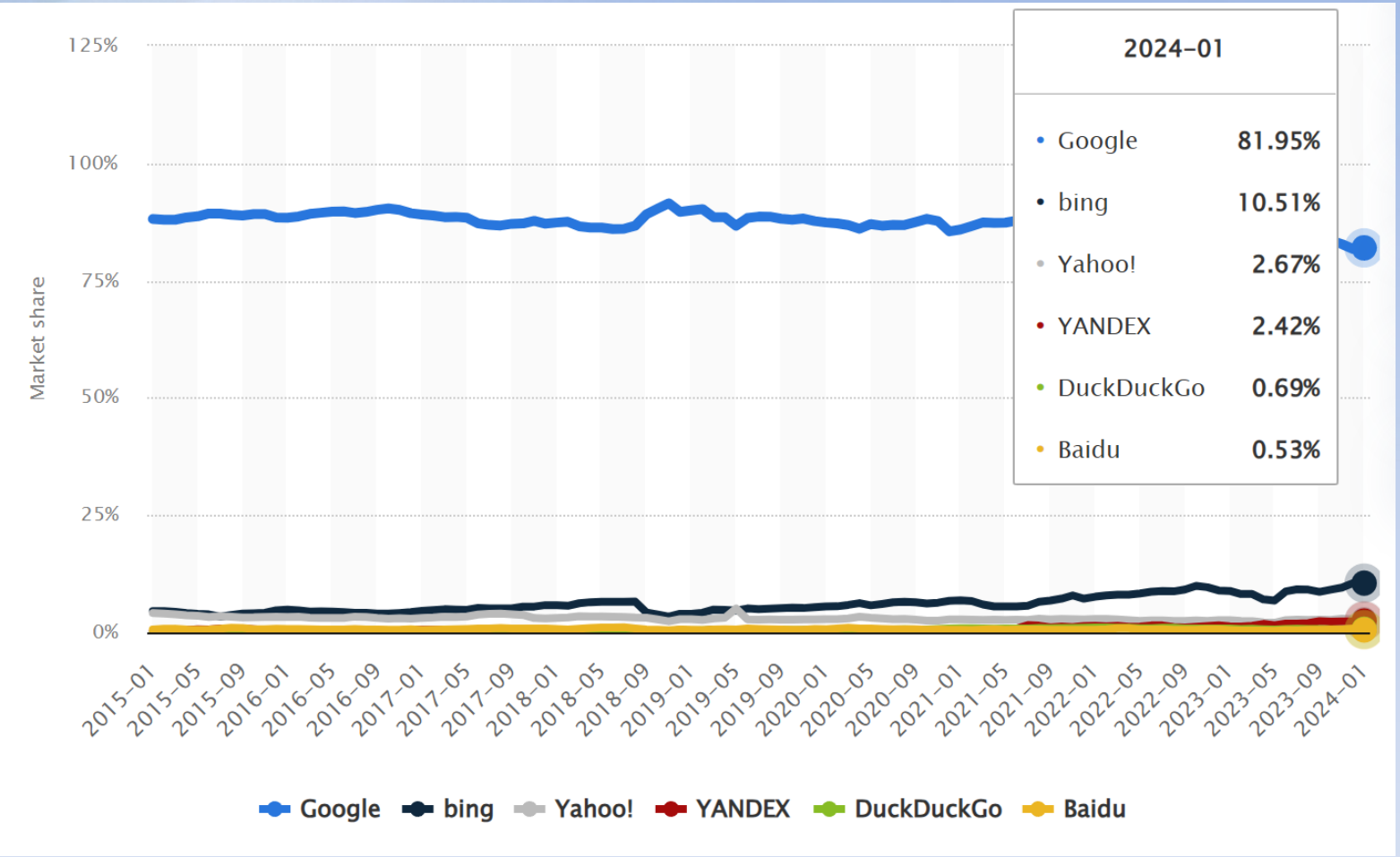
History of search engines

Source: http://en.wikipedia.org/wiki/Search_engines

Year	Name	Event
1993	W3Catalog	Start
	Aliweb	Start
	JumpStation	Start
1994	WebCrawler	Start
	Go.com	Start
	Lycos	Start
1995	AltaVista	Start
	Daum	Company start date
	Open Text Web Index	Start
	Magellan	Start
	Excite	Start
	SAPO	Start
	Yahoo!	Start as a catalogue
1996	Dogpile	Start
	Inktomi	Start
	HotBot	Company start date
	Ask Jeeves	Company start date
1997	Northern Light	Start
	Yandex	Start
1998	Google	Start
	MSN Search	Start
1999	AlltheWeb	Start
	GenieKnows	Company start date
	Naver	Start
	Teoma	Company start date
	Vivisimo	Company start date
2000	Baidu	Company start date
	Exalead	Company start date
2002	Inktomi	Purchase
2003	Info.com	Start

Year	Name	Event
2004	Yahoo! Search	Startas a search engine
	A9.com	Closed
	Sogou	Start
2005	Ask.com	Start
	GoodSearch	Start
	SearchMe	Company start date
2006	wikiseek	Company start date
	Quaero	Company start date
	Ask.com	Start
	Live Search	Start as MSN Search
	ChaCha	Start
	Guruji.com	Start
2007	wikiseek	Closed
	Sproose	Closed
	Wikia Search	Start
	Blackle.com	Start
2008	Powerset	Purchased by Microsoft
	Picollator	Closed
	Viewzi	Closed
	Cuil	Start
	Boogami	Start
	LeapFish	Beta Start
	Forestle	Start
	VADLO	Start
	Duck Duck Go	Start
2009	Bing	Start as Live Search
	Yebol	Start beta version
	Mugurdy	Closed
	Goby	Start
2010	Yandex	StartEnglish version
	Cuil	Closed
	Blekko	Start beta version
	Viewzi	Closed
	Yummly	Start

Search engine market - history

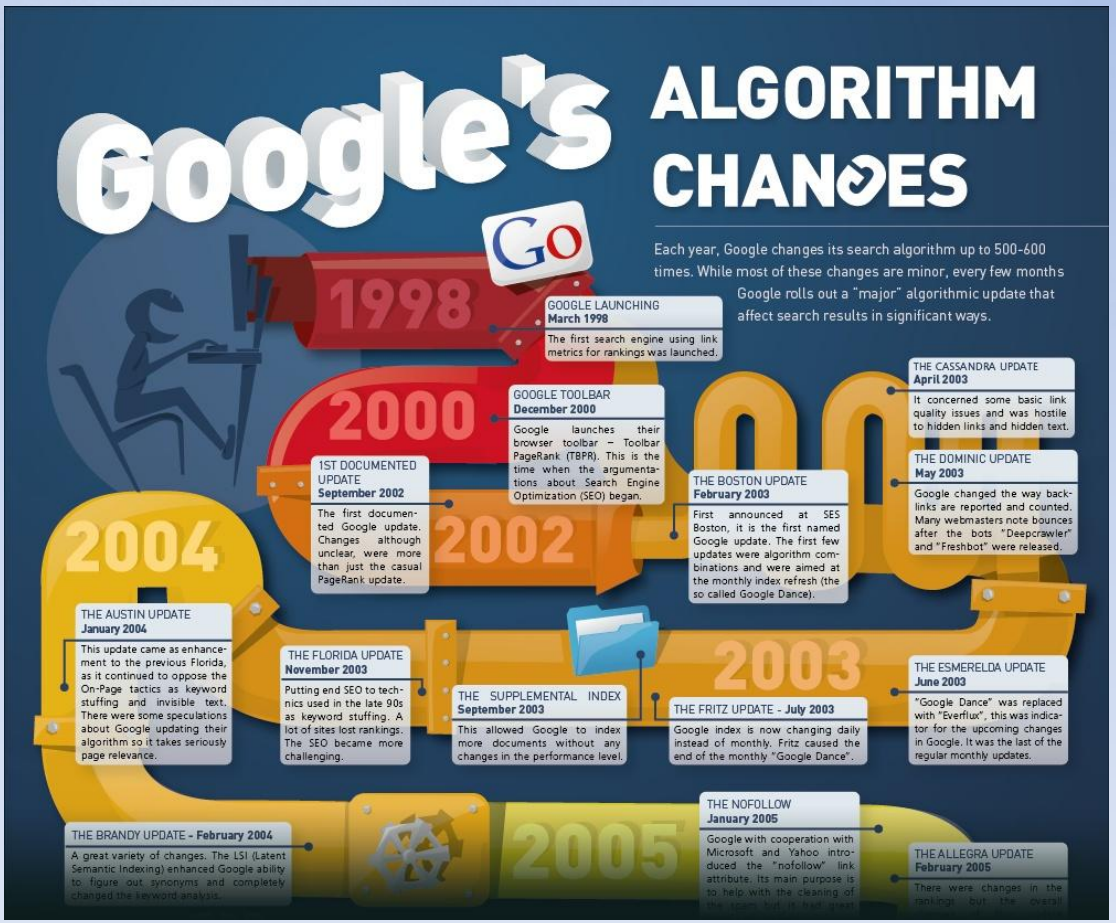


2024-01

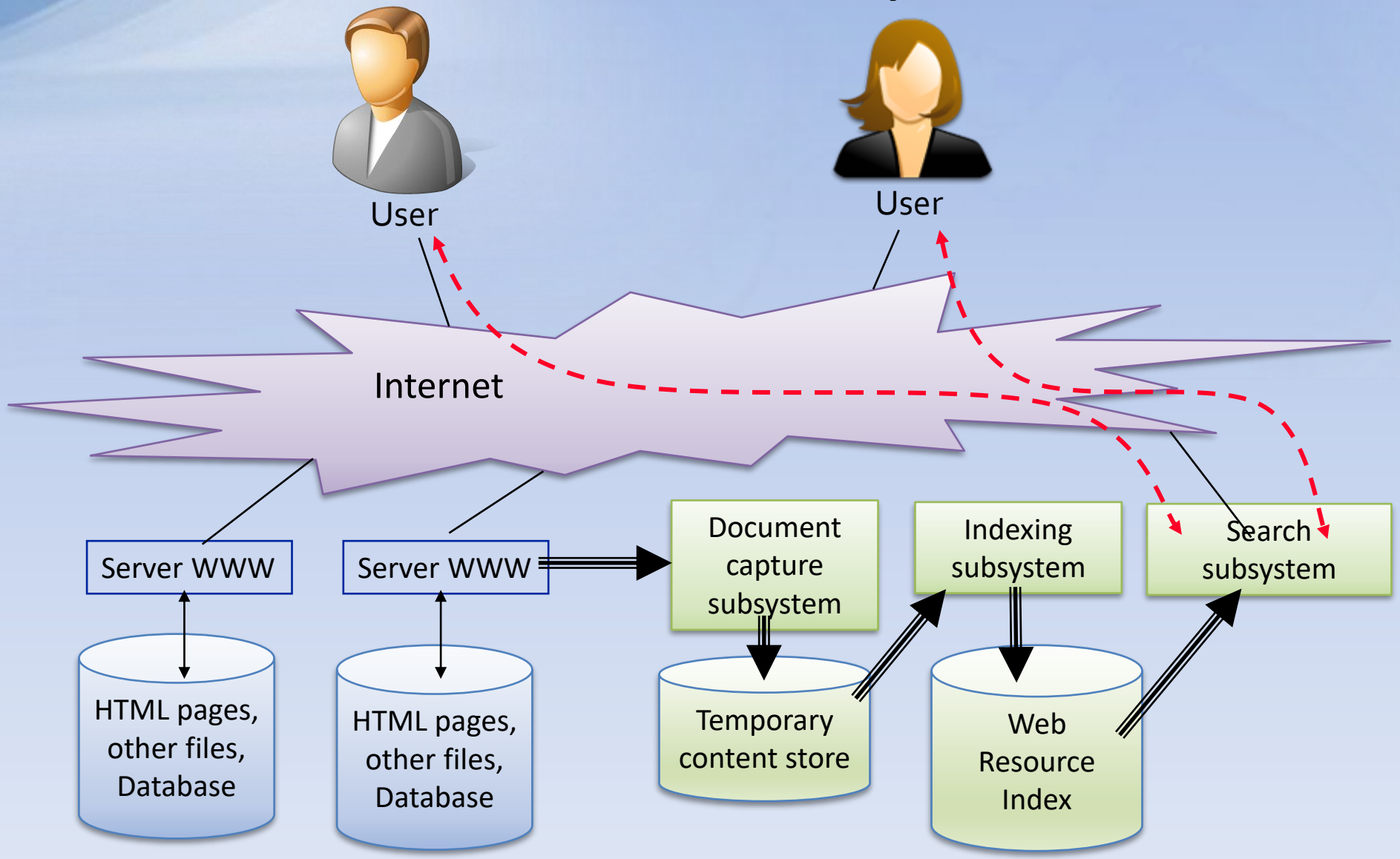
https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

Changes to Google (1998 – 2012)

http://www.seopalbg.com/blog/google-algorithm-changes-1998-2012-infographic



Scheme of the search system



Search engine subsystems

- Document capture subsystem - so called robot, spider or crawler. Circulates the network in a cycle (e.g. daily, weekly) by downloading the available documents to the server. Transparent Closure: If you download an X document, then it downloads all the documents whose URLs are inside the X document, by passing the already downloaded documents.
- Indexing subsystem: After picking up a portion of the document, selects the words from the document and insert them into the central index, along with the corresponding URL and possibly the context.
  - The indexing process can be supported by people (categorization).

Search engine subsystems (2)

- Search subsystem: responds to user / user queries and extracts from the index as HTML pages / pages that are sent to the query.
  - The order of the index entries is of great importance.
  - There are special methods (Google, the "hubs" method) that calculate the "pertinency" of an index item and rank the items in the order of decreasing relevancy.
  - A commercial key is also used (link to search engine sponsor at the beginning of link submission).

Noise, loss of information, relevancy, pertinency

- These are measurable features that quantify the quality of the search result.
- Information noise: Unwanted information, unnecessary, provided due to inaccurate query and / or inaccurate search engine.
- Information loss: Desirable information not provided due to inaccurate query and / or inaccurate search engine mechanism.

Noise, loss of information, relevancy, pertinency (2)

- Relevancy: Determines the extent to which the searched information formally matches the query.
- Pertinency: determines the extent to which the information matches the current needs of the user.
  - Information may be relevant, but for example banal or not accurate.
  - Of course pertinency is far more important than relevancy.

The user's search language

- It would be better to call it a "metaphor for search," because they are increasingly not languages, but graphical metaphors that favor the natural behavior of the user during the search.
- The formal search language is not pleasant to the user. The more formal, the more unfriendly. A more formalized language does not necessarily mean increased pertinency. Only pertinency is important to the user.
- The lack of a formalized search language is not pleasant to the user because it promotes information noise and lowers the level of pertinency.

The user's search language (2)

- Language must meet engineering performance criteria:
  - Time and cost of creating the whole system, cost of operation
  - Quality of service for profit (usually financial)
- Can the golden mean be found?

Smart search engines (AI)

- Dream and subject matter of many academic centers.
  - Is it not a daring dream, in view of the scale of the Web and the necessary engineering compromises?
- Web is a very special database
  - Objects have a highly heterogeneous structure that is not defined; even if it is defined, objects may have a wrong structure.

Smart search engines (2)

- Web is a very special database – *continued*
  - The number of objects and their types is constantly growing
  - Objects are differentiated by topic and meaning, information can be formally and materially incoherent.
  - Objects form a special semantic web through hyperlinks. The semantic meaning of hyperlinks can be arbitrary.
  - The meaning of an object can be determined by its associated objects

Smart search engines (3)

- At the same time the potential user of the Web is very demanding:
  - He/she is not and does not want to be a computer scientist; informative jargon is perceived as unfriendly gibberish.
  - He/she hates reading any manual, "help" and advice.
  - He/she quickly learns some patterns of behavior when working with the Internet and very reluctantly modifies or changes them.

Smart seach engines (4)

- Potential user – *continued*
  - He/she knows almost perfectly (from the position of IT specialist) his/her field of professional activity and expects professional answers to his/her inquiries.
  - ... but does not attach importance to precise, formal questioning.
  - It is expected to help with any difficulty and friendly computer behavior in any situation.
  - He/she does not have the patience to watch hundreds of documents, usually losing patience after 10th document that is incorrect.
  - He/she is keen on getting good results quickly.

Smart search engines (5)

- Some believe that this means the need to integrate some intelligence into the search engines. This can be done, for example, by the following mechanisms:
  - Informing users of the importance of the document, such as the presentation of automatically generated "abstracts", other keywords, the validity of the document.
  - Adaptation (through automatic "learning" of the system, personalization) to individual user preferences.
  - Suggestions for further or alternative directions of exploration.
  - Incorporation of various statistics and measurements about the behavior of the entire population of users to discover certain regularities.

Smart search engines (6)

- It is believed that this requires an automatic "understanding" (?) of the text and the use of "knowledge discovery" technology: Classification and information clustering algorithms.
- Currently, we can observe some attempts to use „inteligent/smart” technologies in the search engines area – *see further*.

Geo-oriented search

- The search engine also stores information about the geographic location of information, such as Poland or even Warsaw,
- Knowing the location of the user (e.g. Poland) you can increase the weight of the results from the services from the same (or similar) location,
- In addition, the native user language can also be considered,
- Enables significant improvements in the quality of results
- It gives you the opportunity to implement a new category of services, e.g. find the nearest pizzeria.

Search using natural language

- It raises huge problems, especially in languages with rich morphology such as Polish or German.
  - Some of these problems are greatly reduced in English.
- The advantage of natural language is that the user knows it (or at least think he/she does) and does not need to learn it.
- It is also flexible, allows you to express any information.

Search using natural language (2)

- However, there are numerous disadvantages of natural language as a means of searching:
  - It is informal and non-formal (especially semantics) at this stage of knowledge. Accordingly, the automatic "understanding" of the text is today only a pseudo-scientific rhetoric (anthropomorphism).
  - The same information can be expressed in any number of ways, which makes it difficult to automatically determine compliance.
  - Information can be differently understood by different people.
  - The same words or sentences can have different meanings depending on the arbitrarily wide context and the associations that arise in the mind of the recipient.

And, Or, Not

- Search through simple Bool algebra. See Google Advanced Search.
  - A document index is a collection where each item is labeled with a set of keywords
  - So for every keyword we have a subset of the index entries - those that bear that word.
  - We can create expressions consisting of keywords, operators AND, OR, NOT and parentheses.
  - AND operator on two subexpressions means theoretically-multitude intersection of the corresponding subsets of the index entries.
  - Operator OR acting on two subtractions is the theoretically – multitude sum of the corresponding subsets of the index entries.
  - A NOT operator acting on the subtraction is a collection created by subtracting from the total index of the subtraction position.

Phrases

- Quite often word sequences have a specific meaning, specific to this sequence.
- The user can search for information based on a frequently heard sequence of words.
- In Polish the additional difficulty is that the order of words in such a phrase can be changed:
  - "Materialized Perspective" <==> "Perspective Materialized"

Phrases (2)

- The consequence is the need to introduce not only individual words into the index, but also their frequent sequences.
  - The problem is how to identify such phrases, how to arrange an index automatically for such phrases and how to use this index
  - You can find some heuristic rules that allow you to treat word sequences as a single search item.
  - In this case, both the phrase and its constituents are index entries.

Inflection

- Indicates the word variation:
  - the word „cat”- „kot”, „koty”, kotu”, „kotem”, „kotami”,...
  - The word „green” - „zielony”, „zielono”, „zazielenić”, „zazieleniony”, „zielone”, „zieleń”,...
- This is so far the most important problem in the construction of search engines, especially in Polish, where the inflection is very rich.
- In English-language search engines, the problem is smaller and in many cases it can be reduced to indexing documents and preprocessing querystring -s or -es.

Inflection (2)

- In the Polish language, the use of a similar method consisting in truncation of several characters from the end and / or from the beginning with simple formal rules leads to significant (unacceptable) information noise, which in an anecdotal way will make our search engine funny in the eyes of users.
- This problem can be partially solved with special dictionaries. Unfortunately, they have a very large volume.

Paradigmatics

- It means semantic dependencies between words or phrases that are independent of their use in the text.
  - The most common type of compound is the termination of the semantic domain of words, e.g. "agricultural machine" 🡐 "tractor". In the index there is a "tractor" and in the user's question is "agricultural machine". How does our search engine associate these terms?
  - The second important connection of this kind is the association, the relationship between the concept and the specific object, which means: "president" 🡐 „Baiden".

Paradigmatics (2)

- *Continued*
  - The third important link is synonym, including abbreviations and acronyms, e.g. "XML" 🡐 🡒 "extended markup language", "tractor" 🡐 🡒 „truck tractor", "database view" 🡐 🡒 "view", .. .
  - There are many other such compounds, in particular the part-whole, e.g. "airplane" 🡐 "propeller" and any combination (see also), eg "Jurek Owsiak" see also "charity".
- A very costly way of mastering the paradigm is to manually construct properly organized dictionaries (so called thesauruses).

Popular spelling mistakes

- Both text makers and search engine users are mistaken.
  - This fact did not get to the developers of information systems whose models were idealistic for years - assumed the correctness of the indexed texts and the faultlessness of the queries.
  - In systems where the user is the only criterion of success, failure to include potential errors is a business failure.

Popular spelling mistakes (2)

- Spelling or grammatical errors specific to a particular language.
- There are also other common mistakes, such as the Czech error (called the "Polish error" in the Czech Republic), skipping letters, etc.

Engineering / business efficiency issue

- The academic world tends to have idealistic treatment of the previously mentioned problems.
- Brute force: Solve the problem as it occurs, in isolation from other problems.
  - For example, many centers have come up with the construction of their own thesauruses to master the problem of paradigm; then, after a few years, these thesauruses were fed with waste paper because of being obsolete.

Engineering / business efficiency issue (2)

- This caused the commercial world to distrust the results produced by the academic world.
  - In practice, it ignores these results and invents its own solutions.
  - These solutions are often put under the great question mark of the aforementioned classic concepts in the field of information search.
- Engineering / business efficiency is determined by the user's satisfaction and the success of the project.
  - Other criteria are secondary.

Solutions to improve search

- Storing previous interests (search criteria) of a user.
- Tracking of visited web pages (requires the installation of dedicated client software)
- Linking to the Internet with a local computer search (MS Windows Desktop Search, Google Desktop Search)
- Assignment of weights to individual words; Popular words have low coefficients
- *More on Google's discussion*

User privacy

- Many search tools collect different user information
- In some cases, it may violate their privacy
- Finding the „golden mean” is not easy.
- Interesting article: *Robert Heaton: How does online tracking actually work?*

User privacy (2)

- Privacy solutions:
  - Customer software (self-hosted), e.g.:
    - SearXNG: a free internet metasearch engine which aggregates results from various search services and databases. Users are neither tracked nor profiled.
  - VPN services, e.g.:
    - Proton VPN,
    - NordVPN,
    - Steganos VPN Online Shield.

Modern „AI” search engines

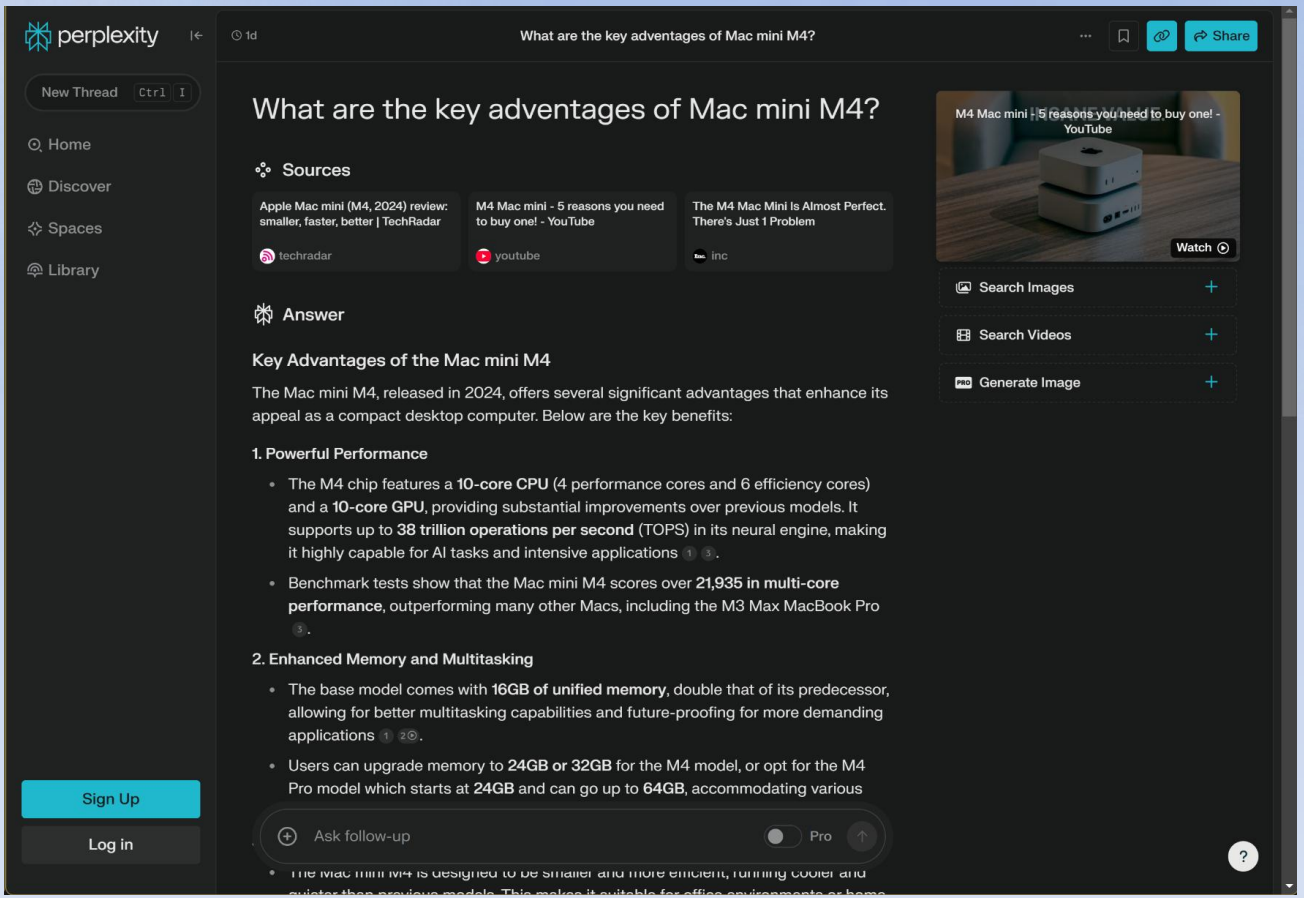
- Real-Time Information: Can browse for the latest data and trends.
- Multi-Source Aggregation: Gathers information from multiple reliable sources.
- Human-Like Explanation: Simplifies complex information into easy-to-understand responses.
- Interactive Clarifications: Users can refine queries for more targeted answers.

Modern „AI” search engines - challenges

- Reliance on Source Availability: Limited by the quality and availability of online sources.
- Potential for Bias: Can inadvertently reflect biases present in source material.
- Not Always Specialized: Best for general or moderately specific questions.
- Ethical Use: Importance of **citing sources** and verifying sensitive information.
- Hallucinations of AI/LLMs.
- Non-determinism (randomness) of AI/LLMs.

„AI” Search engines - **Perplexity**

- A free AI-powered search engine designed to provide real-time, conversational answers to user queries.
- Utilizes advanced language models to interpret questions and **gather information from the web**.
- **Real-Time** Information. Delivers up-to-date data as it pulls information from the internet at the moment of inquiry.



https://www.perplexity.ai

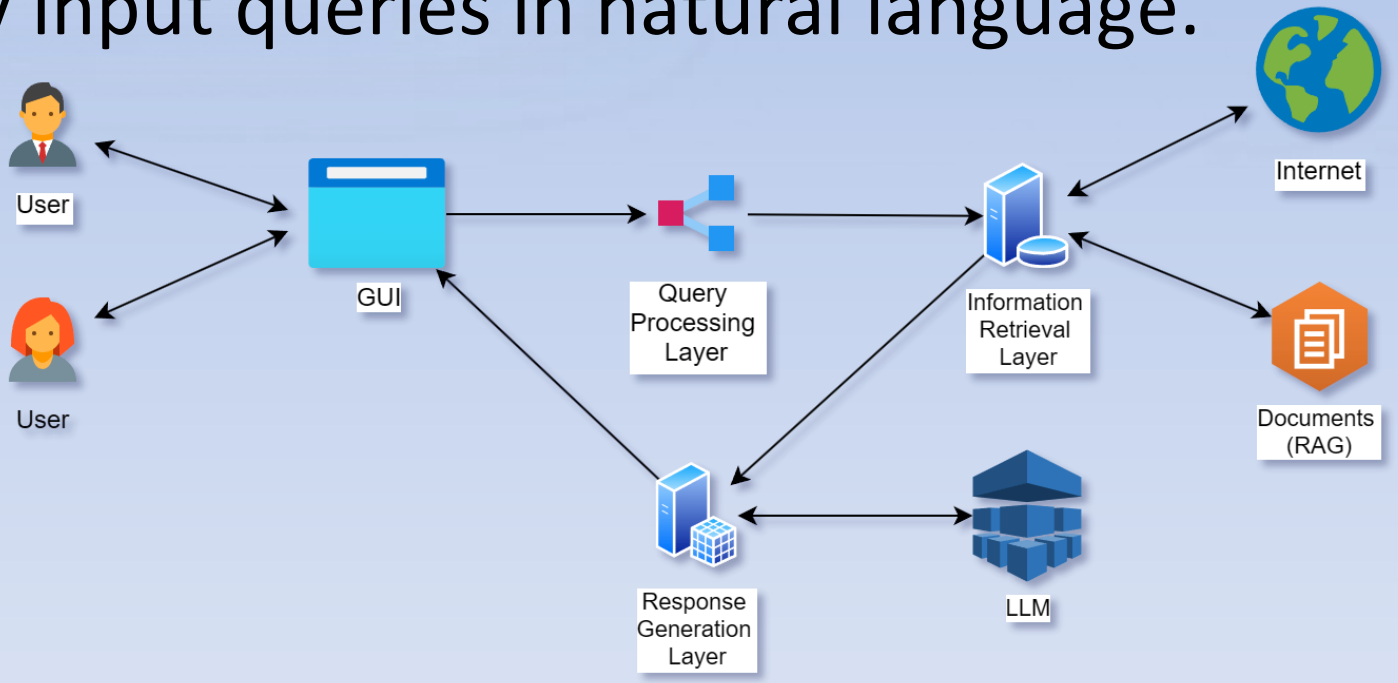
„AI” Search engines – Perplexity (2)

- How Perplexity Works?
  - Understanding the Question. Leverages „AI” to accurately interpret user queries.
  - Searching the Web. Scours the internet for information from authoritative sources, including articles and journals.
  - Summarizing Information. Compiles relevant insights into a clear, concise answer.
  - **Citing Sources**. Provides footnotes linking to original sources for verification and further exploration.

https://www.perplexity.ai

„AI” Search engines – Perplexity (3)

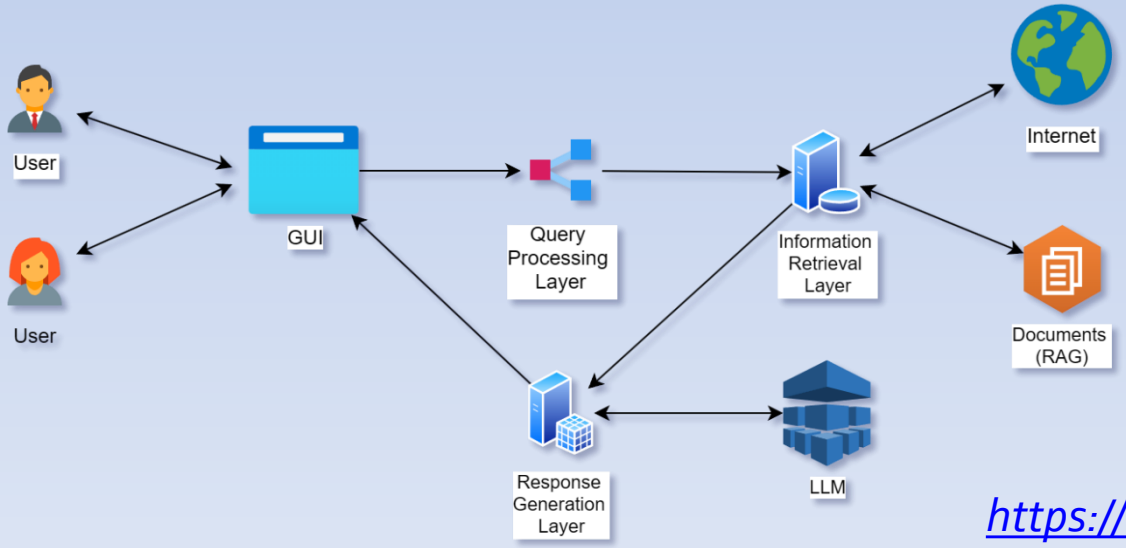
- **GUI**. The user interacts with Perplexity AI through a web interface or mobile app, where they input queries in natural language.



https://www.perplexity.ai

„AI” Search engines – Perplexity (4)

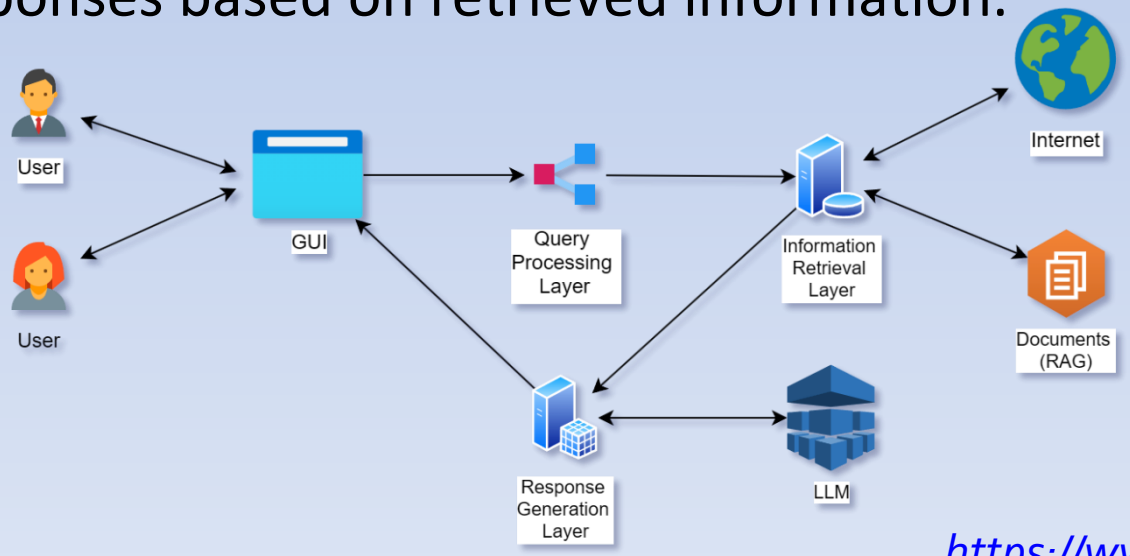
- **Query Processing Layer**.
  - Natural Language Processing (NLP): Utilizes advanced NLP algorithms to interpret and understand user queries. This includes parsing the input to identify intent and context.
  - Contextual Understanding: Employs transformer models to analyze the query's context, ensuring a nuanced understanding of user needs.



https://www.perplexity.ai

„AI” Search engines – Perplexity (5)

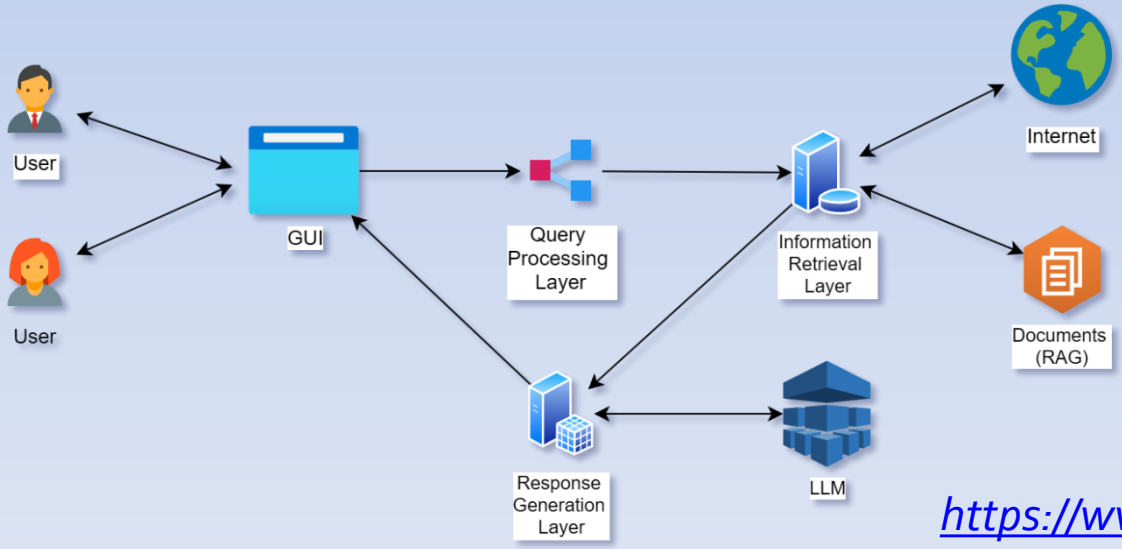
- **Information Retrieval Layer**
  - Web Crawling and Indexing: Searches the internet for relevant data from various sources, including articles, research papers, and other factual content.
  - Retrieval-Augmented Generation (RAG): Combines retrieval of documents with generative capabilities to craft responses based on retrieved information.



https://www.perplexity.ai

„AI” Search engines – Perplexity (6)

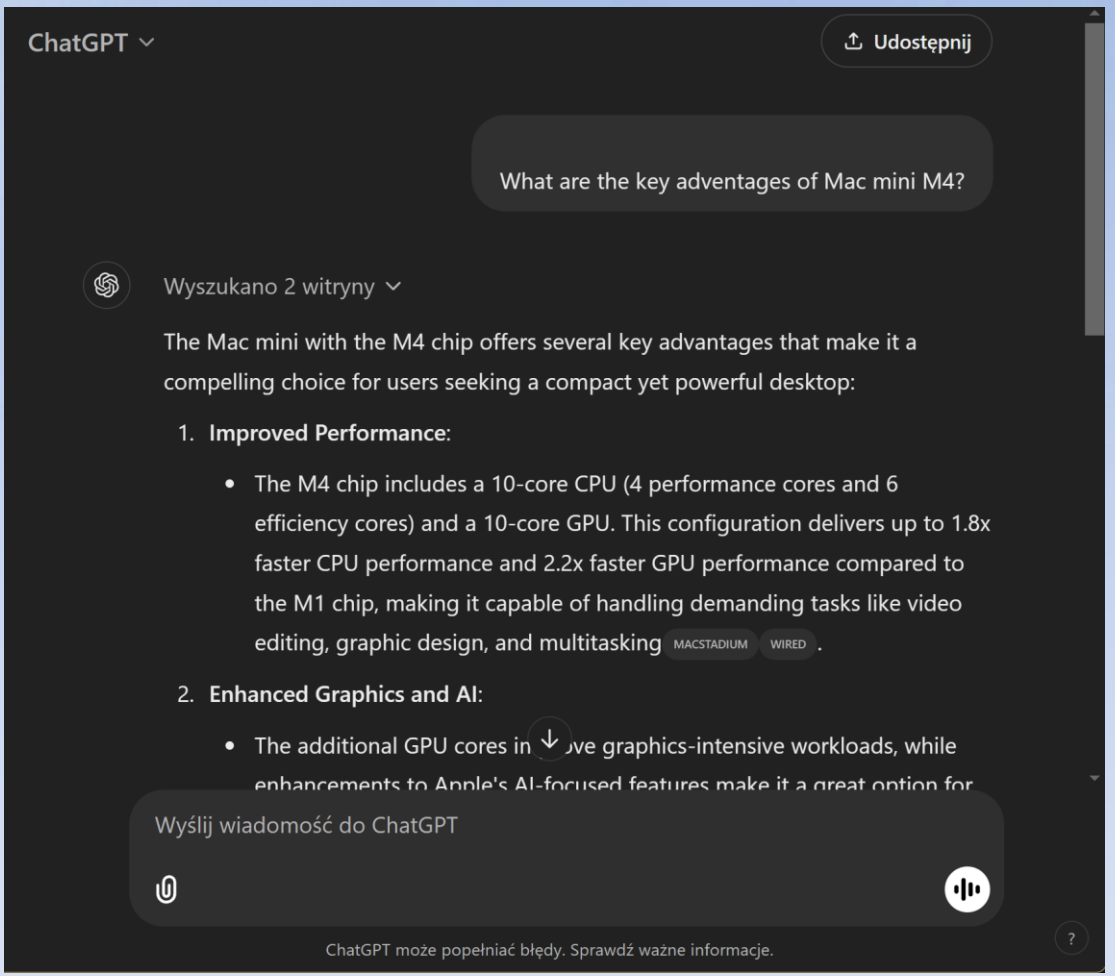
- **Response Generation Layer**
  - Language Model Integration (LLM): Uses large language models (like GPT-4) to generate concise answers based on the retrieved data. This model is fine-tuned for tasks such as summarization and citation referencing.
  - Citing Sources: Each response includes citations from the sources used, allowing users to verify information.



https://www.perplexity.ai

„AI” Search engines – **ChatGPT online**

- ChatGPT is an „AI” language model designed for conversational interactions.
- When functioning as an internet search engine, it uses an integrated browsing tool.
- It retrieves **real-time information**, synthesizes results, and presents insights in natural language.



https://chatgpt.com/

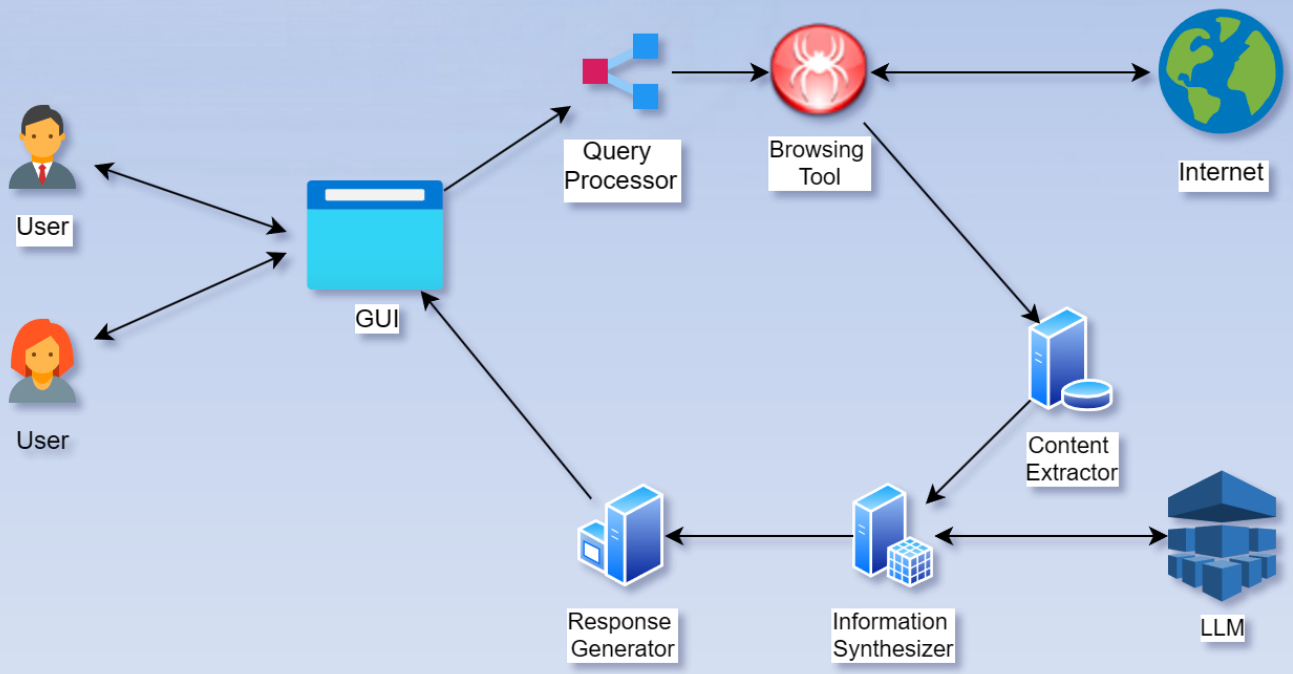
„AI” Search engines – ChatGPT online (2)

- How it works?
  - User Query: A user inputs a question or topic.
  - Search Process:
    - ChatGPT initiates a search using its browsing tool.
    - It identifies relevant sources and collects information.
  - Analysis & Synthesis:
    - Extracts key details from multiple sources.
    - Generates a coherent and informative response.
  - Response Delivery: Delivers the synthesized answer in conversational language.

https://chatgpt.com/

„AI” Search engines – ChatGPT online (3)

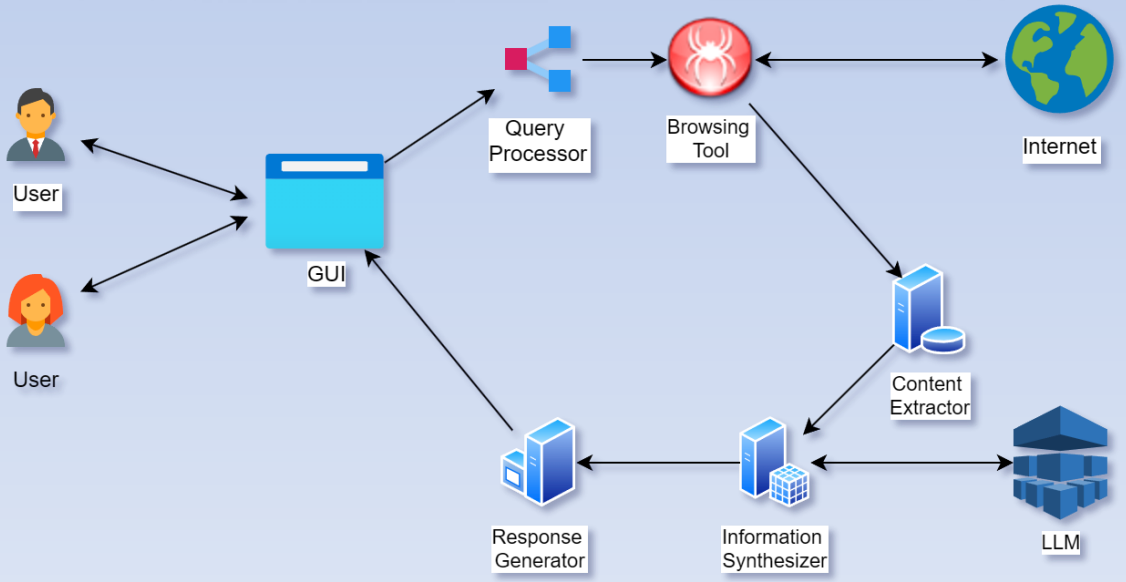
- **GUI**. Users interact through a chat interface, inputting queries and receiving answers.



https://chatgpt.com/

„AI" Search engines – ChatGPT online (4)

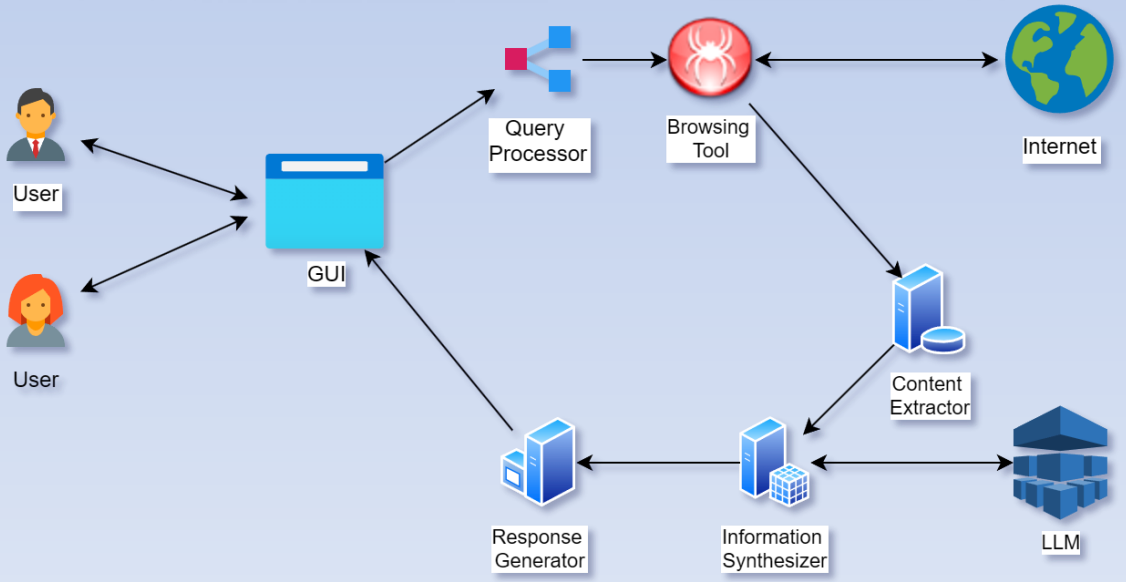
- **Query Processor**
  - Converts the user's natural language query into a structured format.
  - Triggers the browsing tool for web searches.



https://chatgpt.com/

„AI" Search engines – ChatGPT online (5)

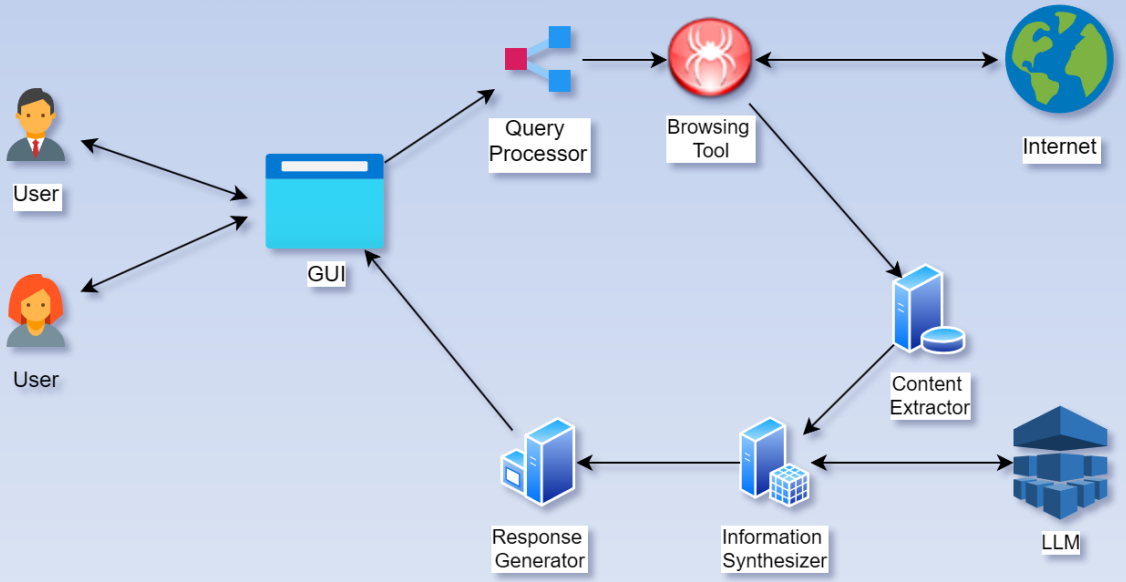
- **Browsing Tool**
  - Executes the search by interacting with search engines.
  - Retrieves top relevant results from the web.



https://chatgpt.com/

„AI” Search engines – ChatGPT online (6)

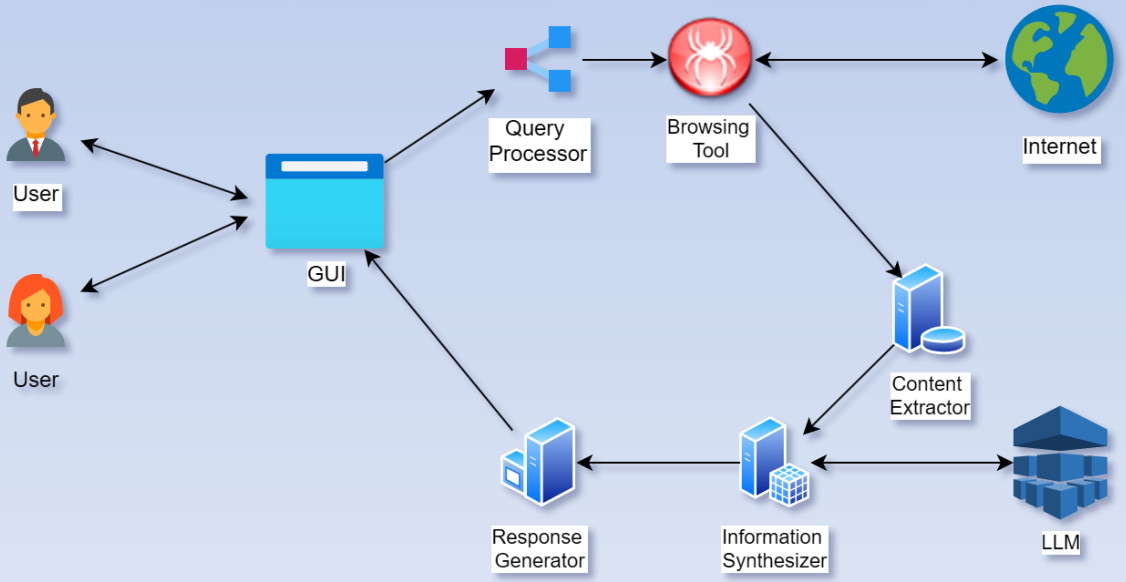
- **Content Extractor**
  - Parses and filters retrieved web pages.
  - Extracts key information based on the user’s query.



https://chatgpt.com/

„AI” Search engines – ChatGPT online (7)

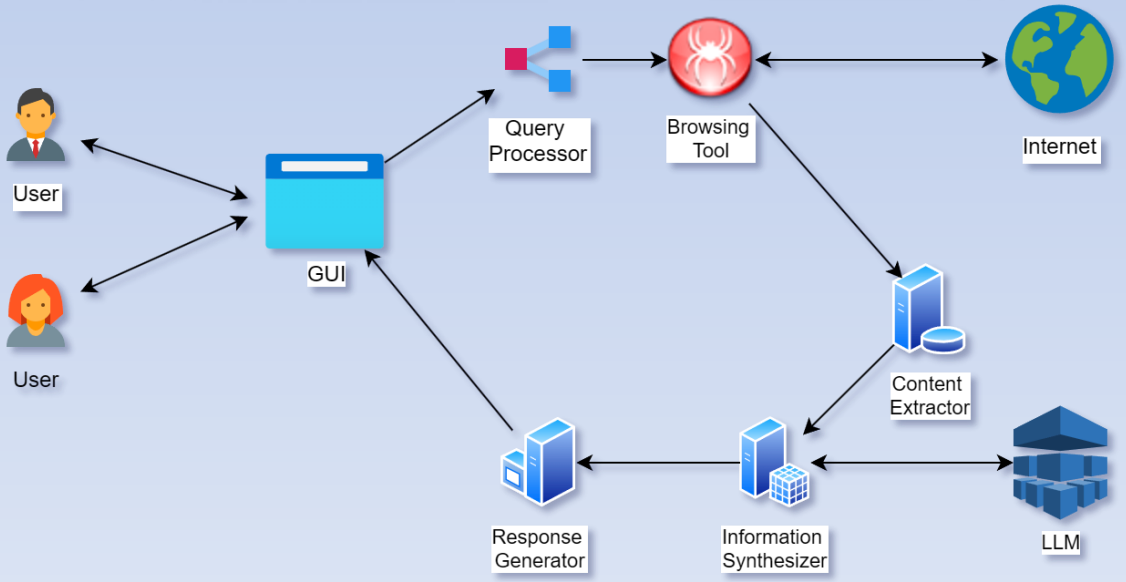
- **Information Synthesizer**
  - Uses natural language processing (NLP) to analyze and merge information.
  - Generates a human-readable response.



https://chatgpt.com/

„AI" Search engines – ChatGPT online (8)

- **Response Generator**
  - Converts synthesized data into a conversational reply.
  - Handles follow-up questions interactively.



https://chatgpt.com/

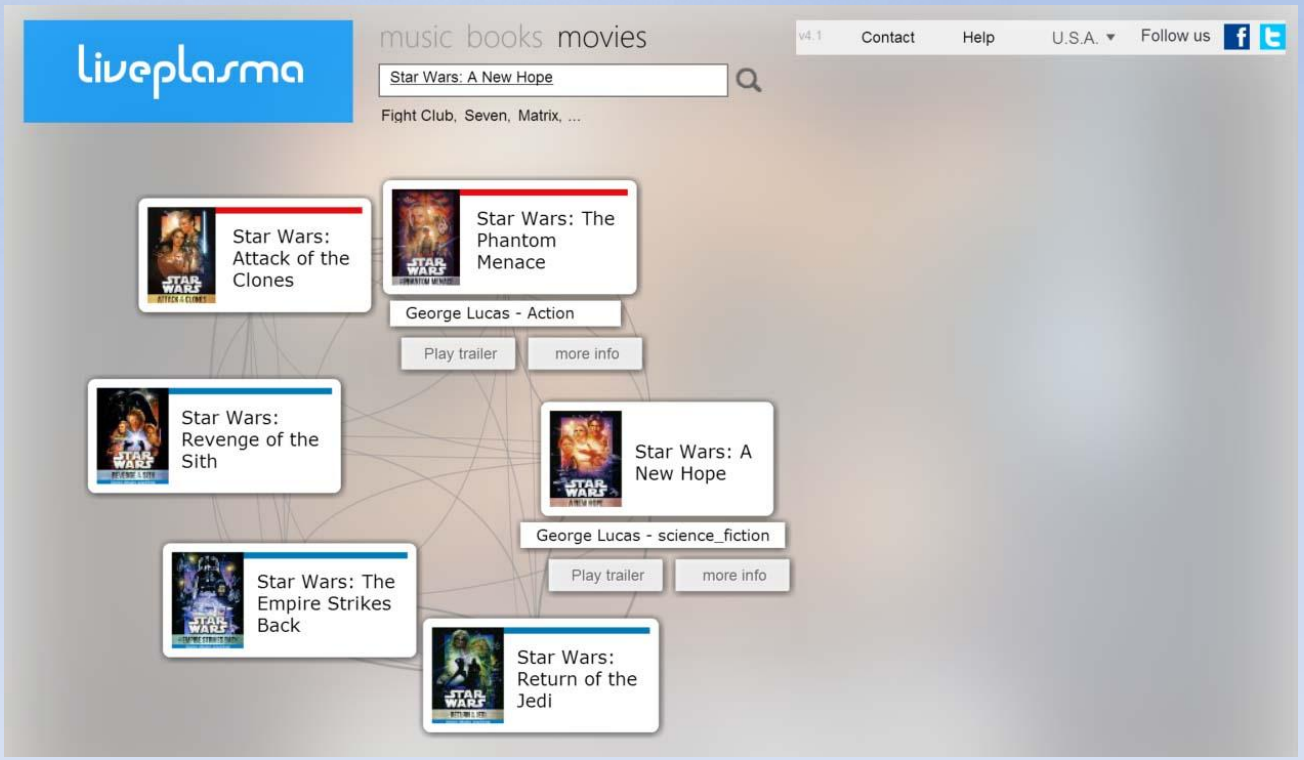
Interesting solutions

- Eyeplorer (http://en.eyeplorer.com/)

Results for: ***Arduino***

Interesting solutions (2)

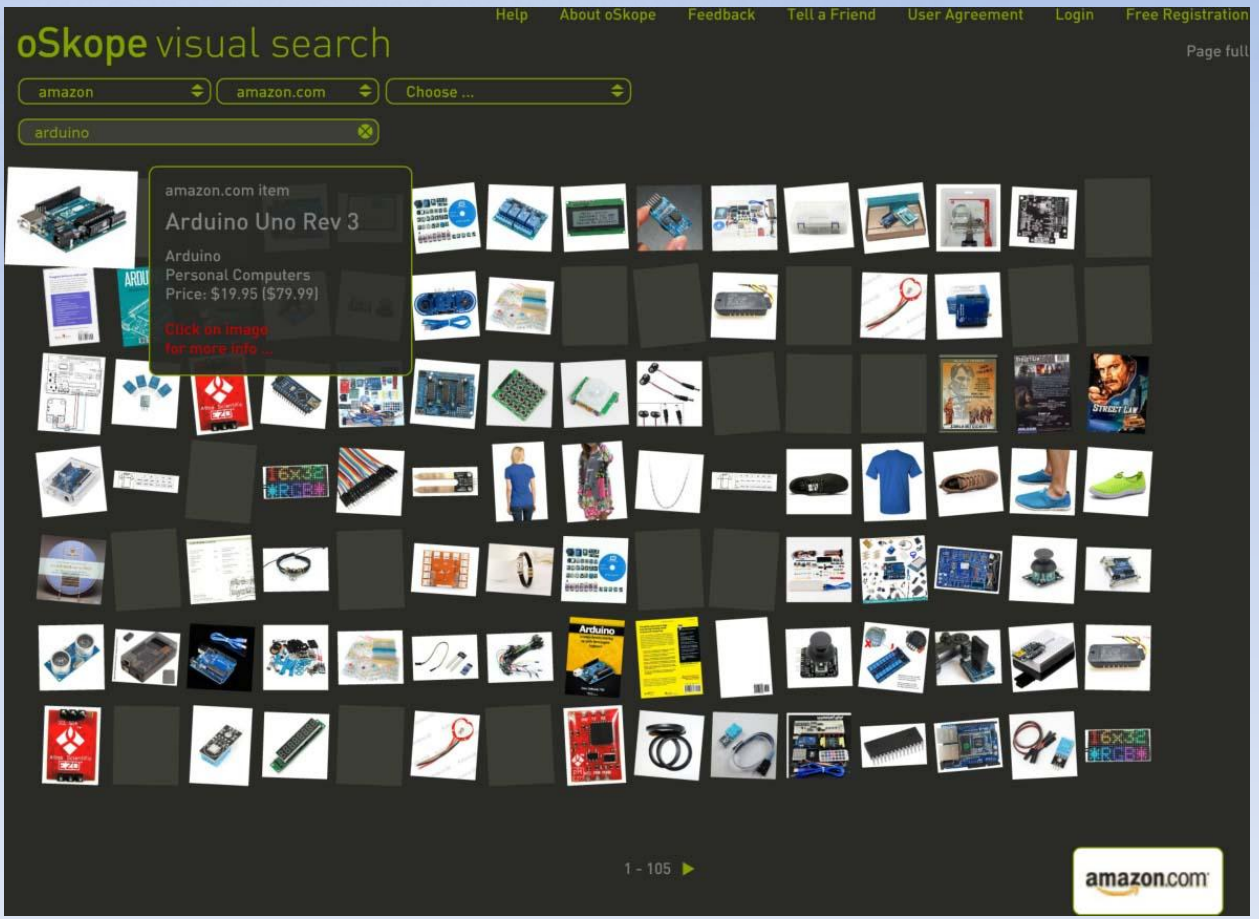
- LivePlasma (http://www.liveplasma.com/)



Results for: ***Star Wars: A New Hope***

Interesting solutions (3)

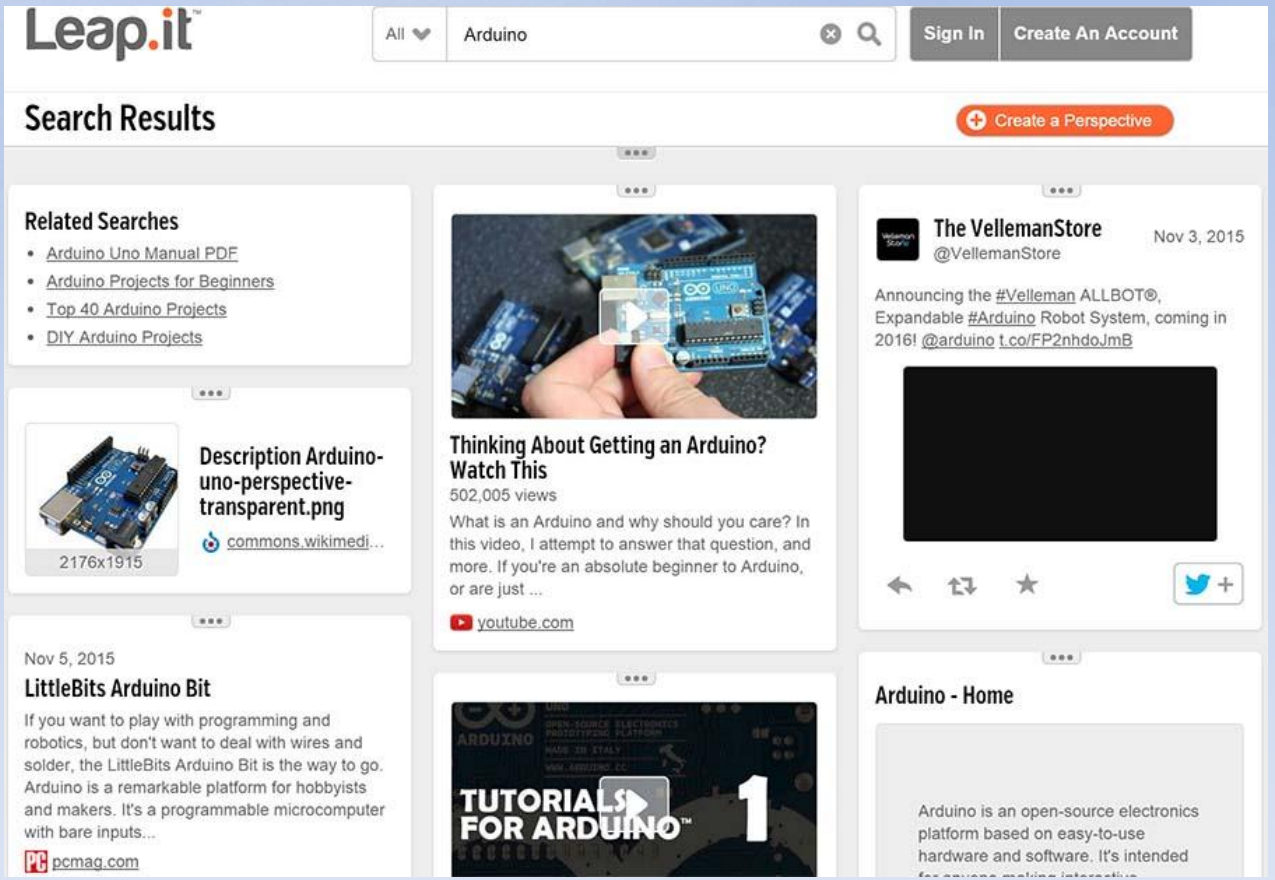
- oSkope (http://www.oskope.com/)



Results for: ***Arduino***

Interesting solutions (4)

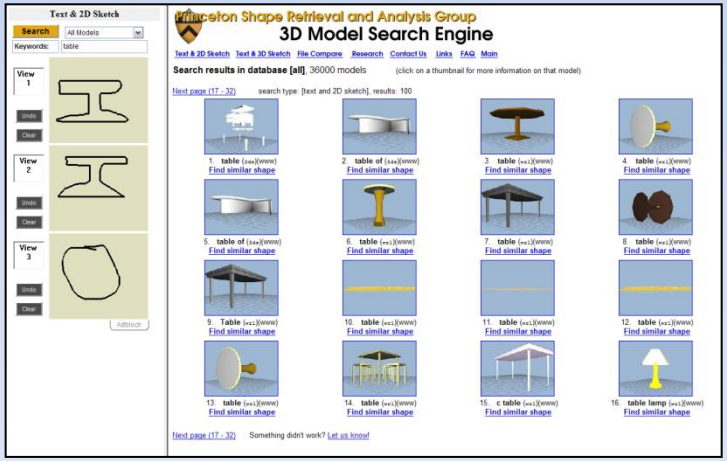
- Leap It (http://leap.it/)



Results for: ***Arduino***

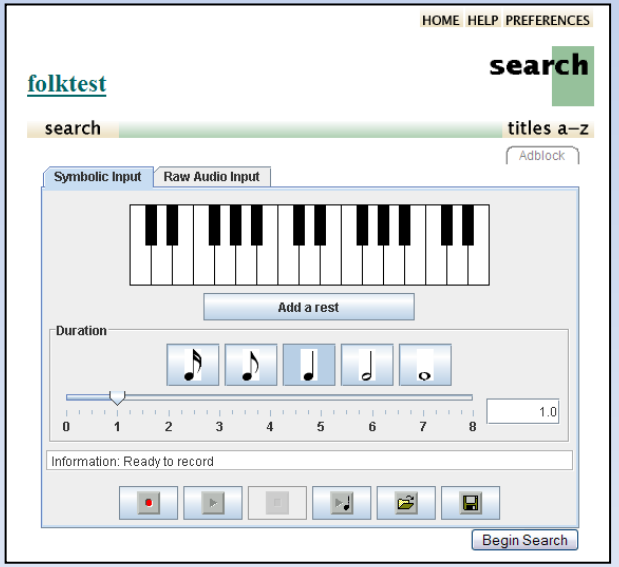
Search for other media

- Searching through other paradigms than text (such as graphics or music), presents many difficulties.
- The Princeton 3D Model Search Engine allows you to search for 2D and 3D models(http://shape.cs.princeton.edu/search.html)
  - Keywords
  - Recognition of shapes based on the user
  - The effects are promising, but still need refining



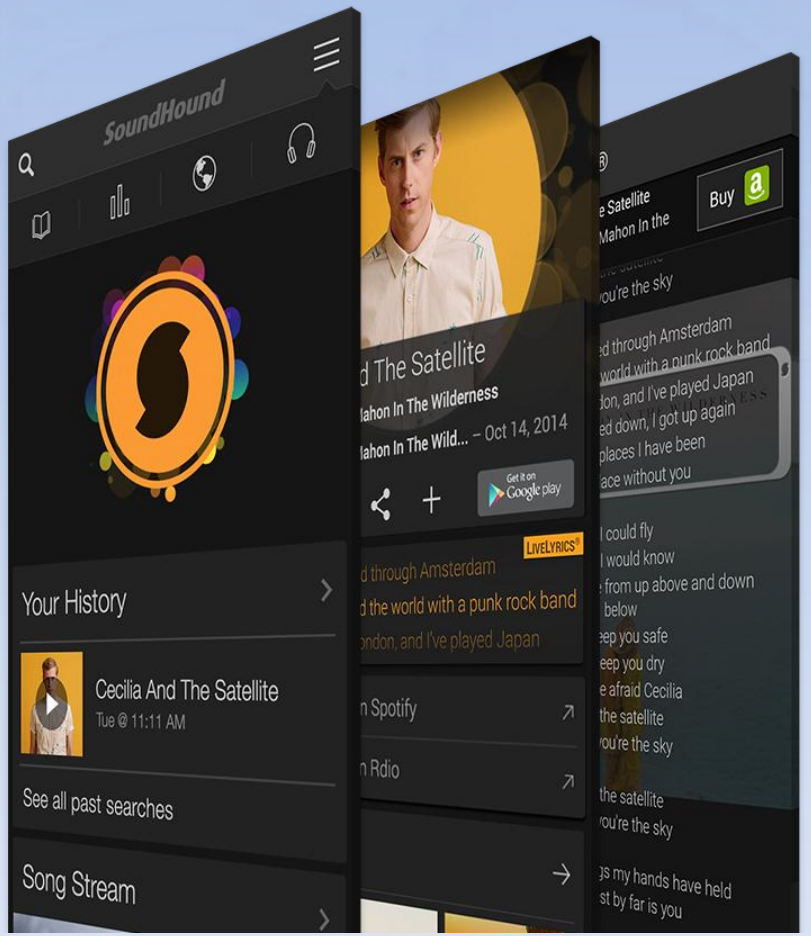
Search for other media (2)

- Music search
  - How to define a query?
    - A musical note or some variant thereof
    - To play a song piece using the keyboard
    - Hum the fragment into the microphone
    - Lyrics
  - Meldex created within New Zealand Digital Library Project (http://www.nzdl.org/musiclib)



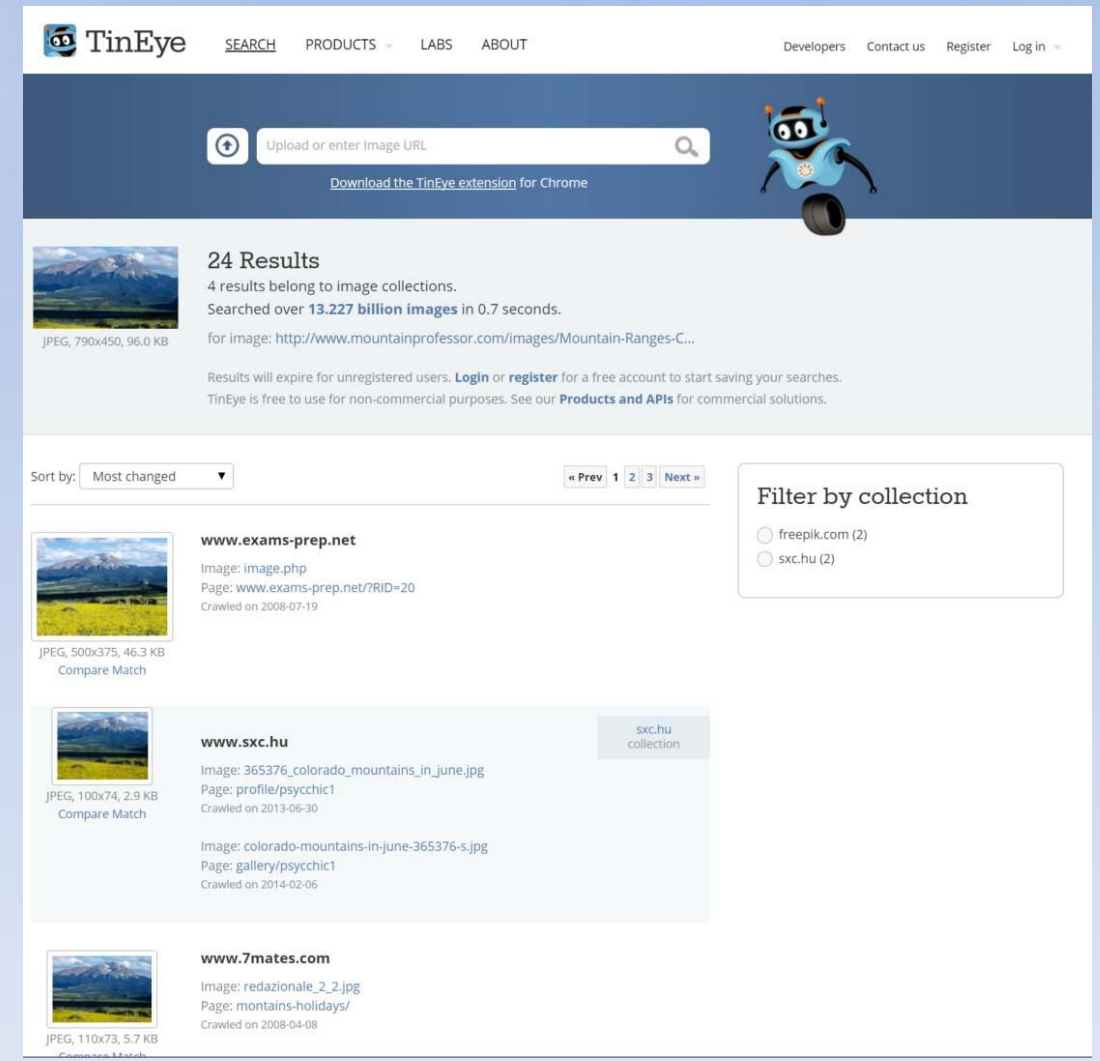
Search for other media (3)

- Music search– continued
  - Mobile application **SoundHound** (http://www.soundhound.com/) automatically recognizes the music played.



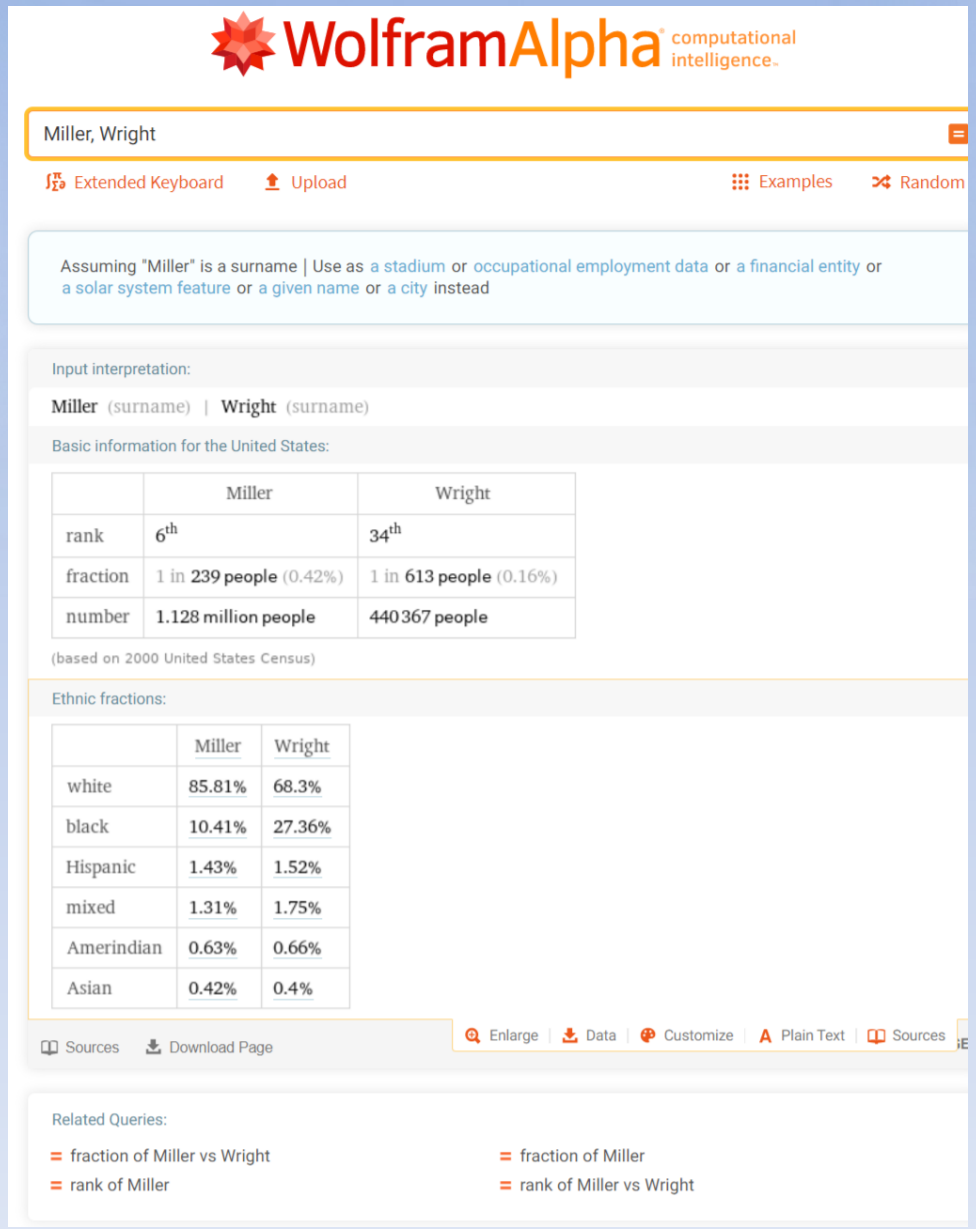
Search for other media (4)

- Image search
  - **TinEye** (http://tineye.com/) recognizes duplicate graphics (even modified)
  - Feature in Google Chrome



Other search engines

- WolframAlpha
  - „math searches”
  - various examples: *https://www.wolframalpha.com/examples/*



Summary

- Searching for information online is a very complex matter.
- Search engines not only make it easy to use the Internet, but they make it possible.
- Today the market is dominated by one company: Google.
- It will be so, until someone else invents new, groundbreaking ways of searching.
- So far, there are no good search solutions using other methods than text.
- The results returned by popular websites are getting better, but you can still wish for more and better ways.