CS8080 – Information Retrieval Techniques

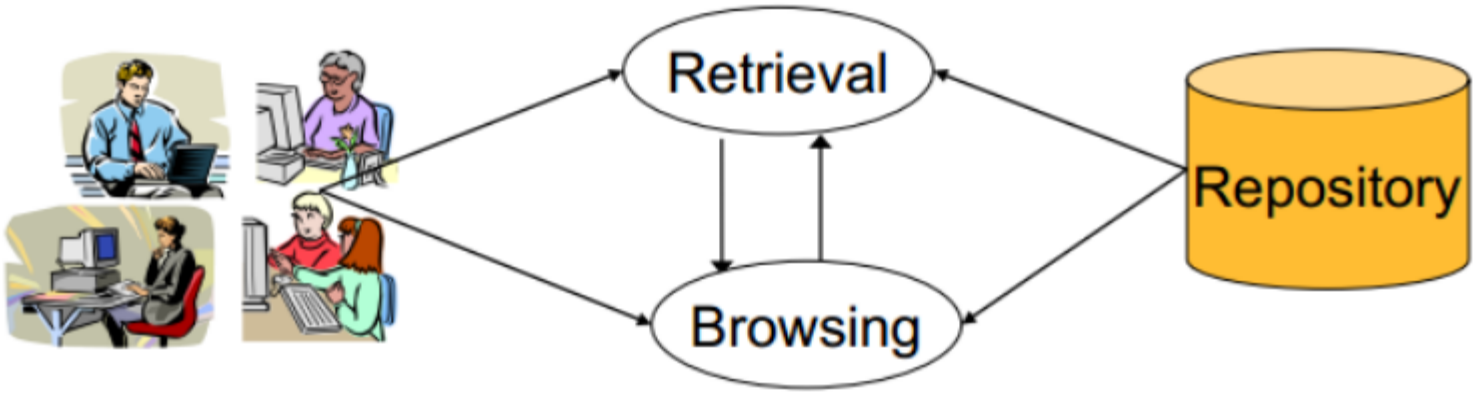
An IR model involves in the following activities

- Document Matching and Ranking
  - It involves finding the relevant documents from the document repository that may contain the answer to the user query .it is also rank them as per its relevance to the user query .

- Two main search paradigms

  Retrieval and Browse

  - Retrieval
    - Search for particular information
    - Usually focused and purposeful
  - Browsing
    - General looking around for information
    - For example: Asia-> Thailand -> Phuket ->Tsunami



IR vs. DBMS

IR	DBMS
Imprecise Semantics	Precise Semantics
Keyword search	SQL
Unstructured data format	Structured data
Read-Mostly. Add docs occasionally	Expect reasonable number of updates
Page through **top** k results	Generate full answer

- Information Retrieval Vs information Extraction

- Information Retrieval:
- Given a set of terms and a set of document terms select only most relevant document (precision) ,and preferably all the relevant ones(recall)

- Information Extraction:
- Extract from the text what the document means.

- Early Developments:

- For centuries indexes were created manually as categorization hierarchies.
- In fact most libraries still use some form of categorical hierarchy to classify their volumes.
- Such hierarchies have usually been conceived by human subjects from the library sciences field.
- More recently ,the advent of modern computers has made possible the construction of large indexes automatically.
- Automatic indexes provide as view of the retrieval problem which is much more related to the system itself than to the user need

- Early Developments:
- An old and popular data structure for faster information retrieval is a collection of selected words or concepts with which are associated pointers to the related information called index.

- In one form or another, indexes are at the core of every modern information retrieval system.
- They provide faster access to the data and allow the query processing task to be speeded up.

- Early Developments:

- Automatic indexes provide as view of the retrieval problem which is much more related to the system itself than to the user need.

- It is important to distinguish between two different views of the IR problem:
- a computer centered one and
- a human centered one

- In the computer centered view ,
- the IR problem consists mainly of building up efficient indexes, processing user queries with high performance, and developing ranking algorithms which improve the quality of the answer set.

- In the Human centered view ,the IR problem consists mainly of studying the behavior of the user, of understanding his main needs, and of determining how such understanding affects the organisation and operation of the retrieval system.

- Information Retrieval in the library:
- Libraries were among the first institutions to adopt IR systems for retrieving information.

- In the first generation, such systems consisted basically allowed searches based on author name and title
- In the second generation increased search functionality was added which allowed searching by subject headings, by keywords and some complex query facilities.
- In the third generation which is currently being deployed the focus is on improved graphical interfaces, electronic forms, hypertext features and open system architectures

- Web and Digital Library
- Three dramatic and fundamental changes have occurred due to the advances in modern computer technology and boom of the web.

they are

- 1.A cheaper access to various sources of information.
- 2.Provide greater access to networks.
- 3.Publishing freedom

- Components of IR System
- Information retrieval locates relevant documents on the basis of user input such as keywords or example documents,
- for example : Find documents containing the words “database systems”.
- the figure shows information retrieval system block diagram. It consists of three components: Query or Documents, IR system and Ranked Results

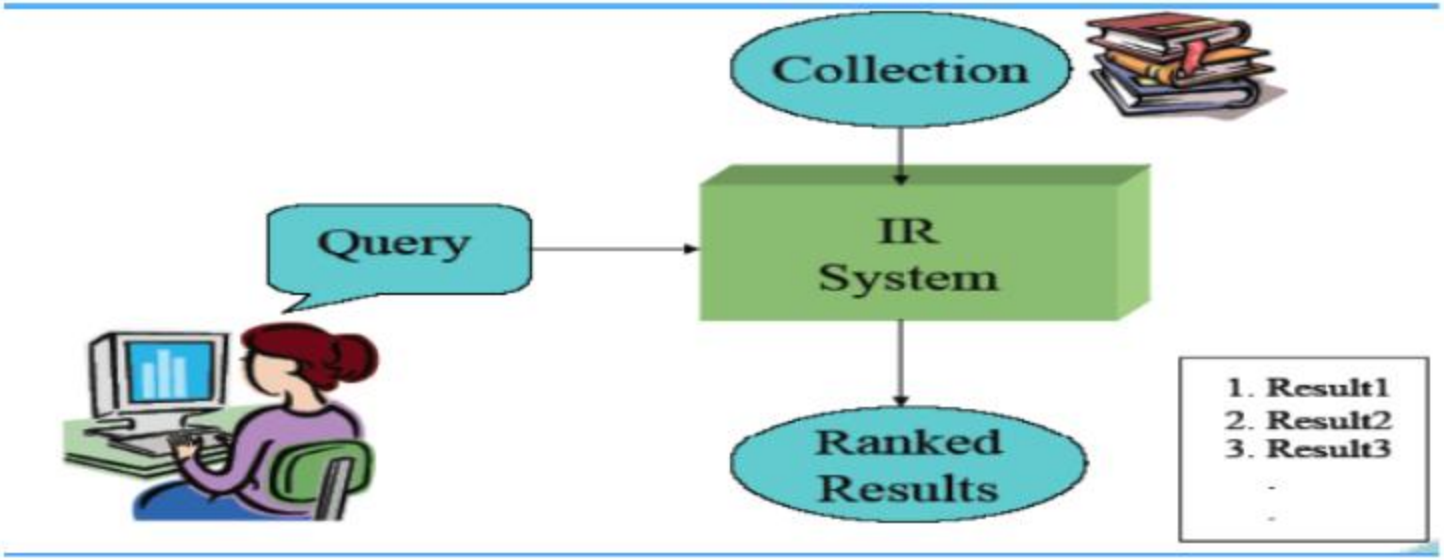
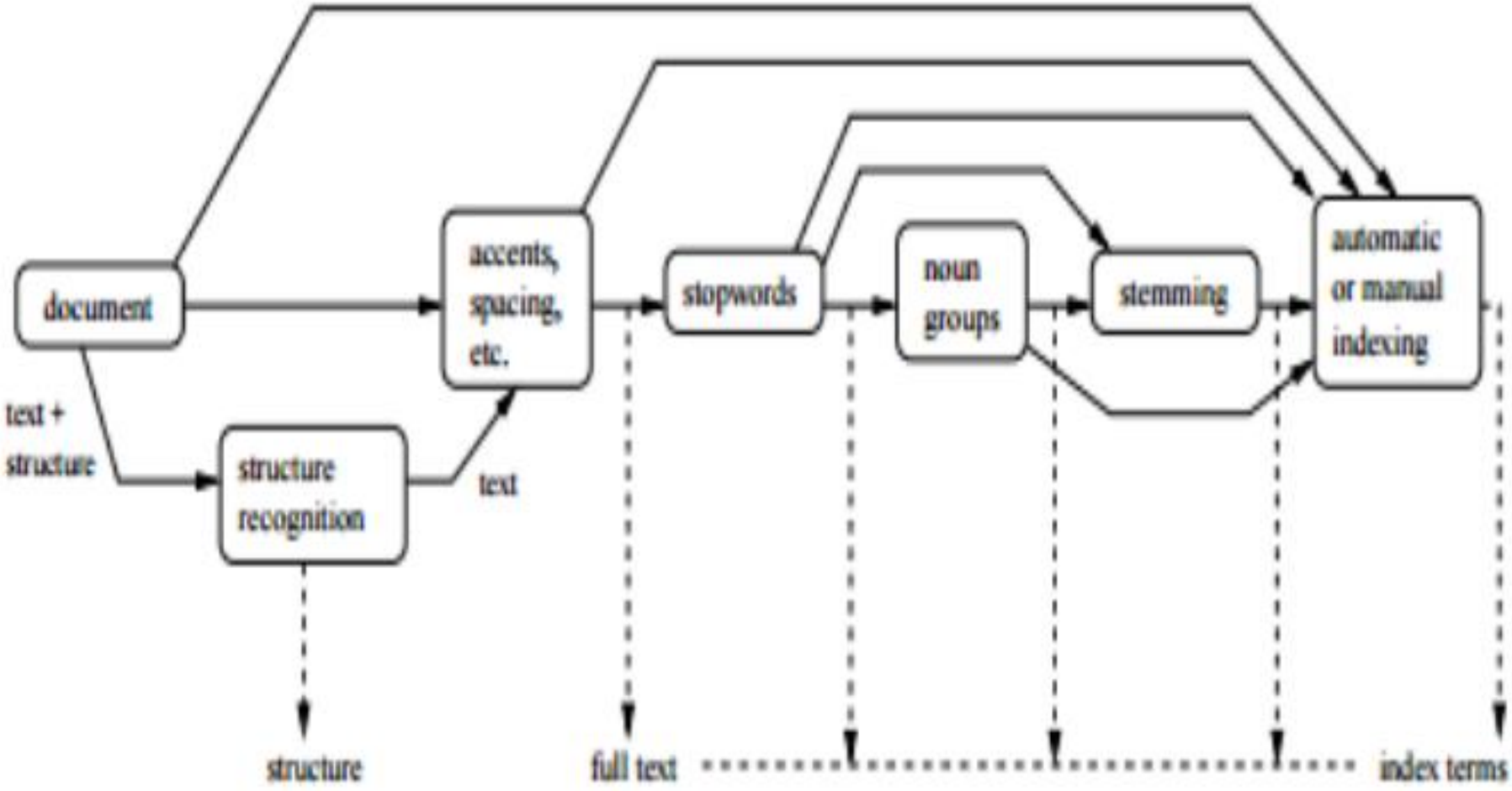


Figure:Block diagram of IR

- Architecture of IR System
- Logical View of the Documents:

- Due to historical reasons, documents in a collection are frequently represented through a set of index terms or keywords.
- Such keywords might be extracted directly from the text of the document or might be specified by a human subject

- No matter whether these representative keywords are derived automatically or generated by a specialist, they provide a logical view of the document


document
text +
structure
structure
recognition
accents,
spacing,
etc.
text
stopwords
noun
groups
stemming
automatic
or manual
indexing
structure
full text
index terms

- With very large collections, however, even modern computers might have to reduce the set of representative keywords.
- This can be accomplished through the elimination of stopwords (such as articles and connectives),
- the use of stemming (which reduces distinct words to their common grammatical root),and
- the identification of noun groups (which eliminates adjectives, adverbs, and verbs).
- These operations are called text operations (transformations).
- Text operations reduce the complexity of the document representation and allow moving the logical view from that of a full text to that of a set of index terms

- The Retrieval Process:
- To describe the retrieval process, we use a simple and generic software architecture as shown in Figure
- First of all, before the retrieval process can even be initiated, it is necessary to done the text database.
- This is usually done by the manager of the database, which species the following:
  - (a) the documents to be used,
  - (b) the operations to be performed on the text, and
  - (c) the text model (i.e., the text structure and what elements can be retrieved).
  - The text operations transform the original documents and generate a logical view of them.

- The Retrieval Process:
- Once the logical view of the documents is defined, the database manager (using the DB Manager Module) builds an index of the text.

- An index is a critical data structure because it allows fast searching over large volumes of data.
- Different index structures might be used, but the most popular one is the inverted index as indicated in Figure .
- The resources (time and storage space) spent on defining the text database and building the index are reduced by querying the retrieval system many times.

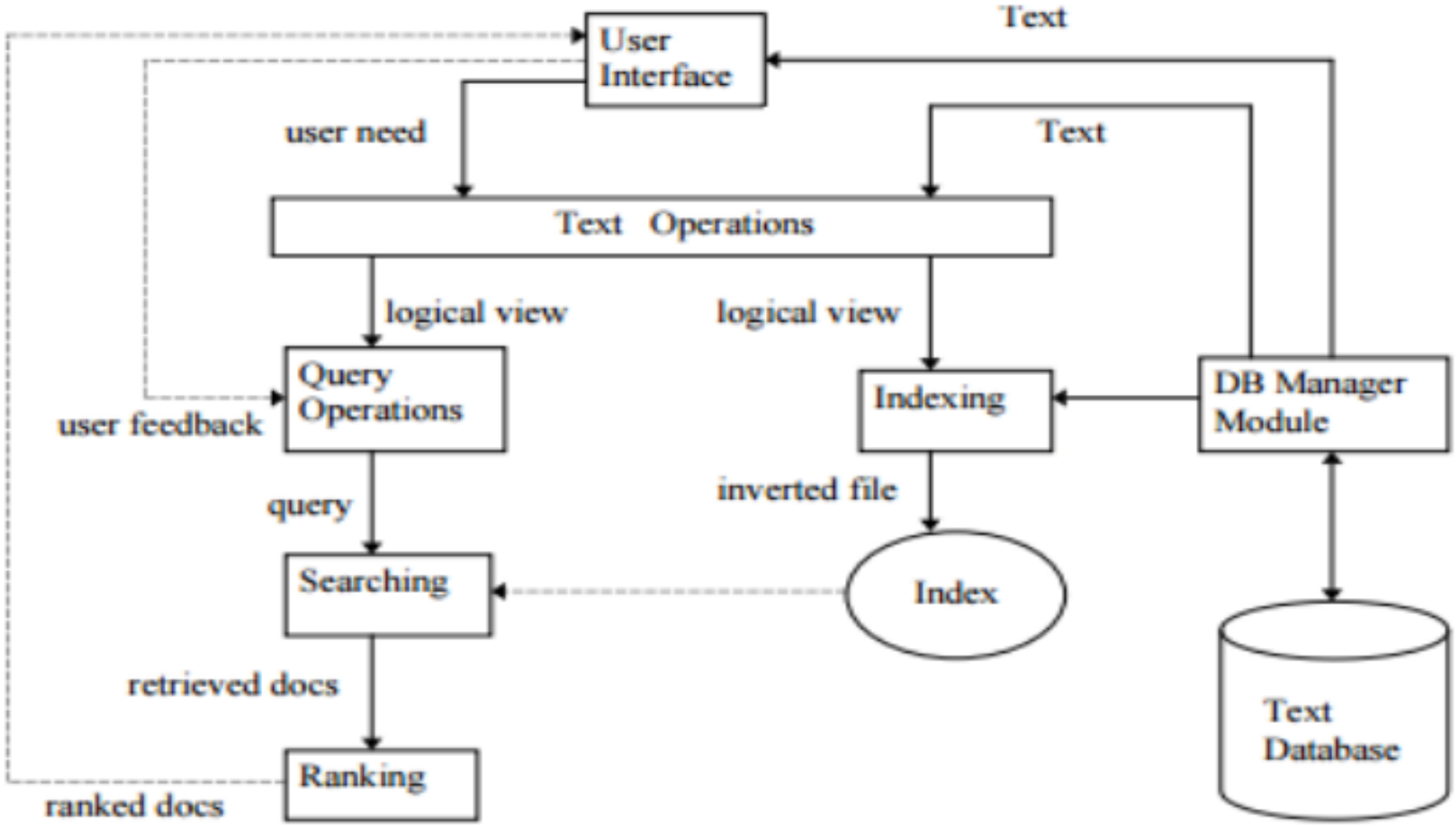


Figure . The process of retrieving information

- Text Operations forms index words (tokens).
  - Stopword removal
  - Stemming

- Indexing constructs an inverted index of word to document pointers.
- Searching retrieves documents that contain a given query token from the inverted index
- Ranking scores all retrieved documents according to a relevance metric

- Issues in IR
- The main objective of an IR system is to retrieve all the items that are relevant to a user query, while retrieving as few non relevant items as possible

Main problems in IR

- Document and Query indexing
  - How to best represent their contents?
- Query evaluation(or retrieval process)
  - To what extent does a document correspond to a query?
- System evaluation
  - How good is a system?
- Are the retrieved documents relevant?(precision)
- Are all the relevant documents retrieved? (recall)

Three Big Issues in IR

- 1.Relevance
- It is the fundamental concept in IR.
- A relevant document contains the information that a person was looking for when she submitted a query to the search engine.
- There are many factors that go into a person's decision as to whether a document is relevant.
- These factors must be taken into account when designing algorithms for comparing text and ranking documents.
- Simply comparing the text of a query with the text of a document and looking for an exact match, as might be done in a database system produces very poor results in terms of relevance.

Three Big Issues in IR

- To address the issue of relevance, retrieval models are used.
- A retrieval model is a formal representation of the process of matching a query and a document.
- It is the basis of the ranking algorithm that is used in a search engine to produce the ranked list of documents.
- A good retrieval model will find documents that are likely to be considered relevant by the person who submitted the query.

- The retrieval models used in IR typically model the statistical properties of text rather than the linguistic structure.
- For example, the ranking algorithms are concerned with the counts of word occurrences than whether the word is a noun or an adjective

Three Big Issues in IR

- Evaluation
- Two of the evaluation measures are precision and recall.
- Precision is the proportion of retrieved documents that are relevant.
- Recall is the proportion of relevant documents that are retrieved.
- Both precision and recall are therefore based on an understanding and measure of relevance.
- Suppose a computer program for recognizing dogs in photographs identifies 8 dogs in a picture containing 10 cats and 12 dogs (the relevant elements). Of the 8 identified as dogs, 5 actually are dogs (true positives), while the other 3 are cats (false positives). 7 dogs were missed (false negatives), and 7 cats were correctly excluded (true negatives).
- So, in this case, precision is "how valid the search results are", and recall is "how complete the results are".

$$\text{Precision} = \frac{\text{Relevant documents} \cap \text{Retrieved documents}}{\text{Retrieved documents}}$$

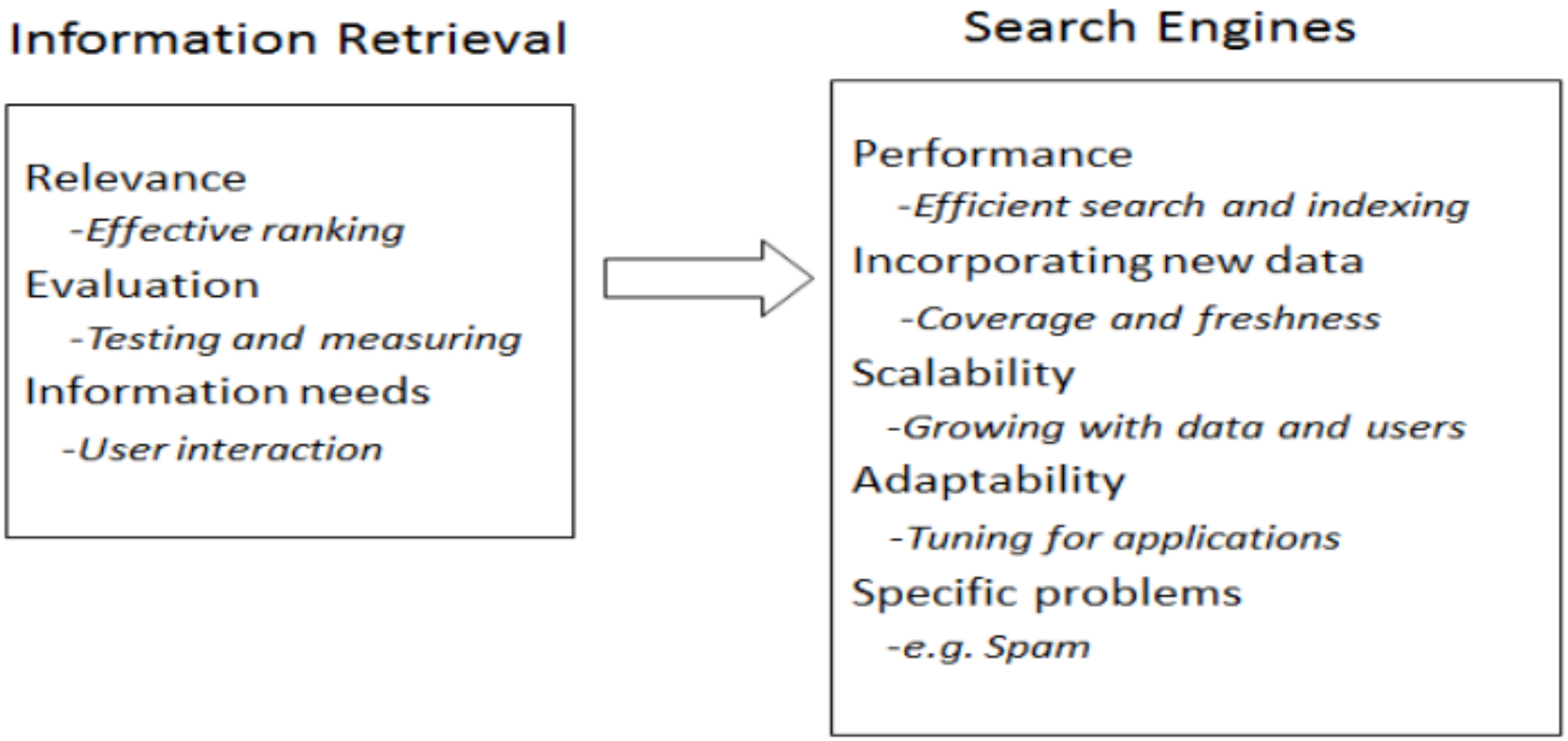
$$\text{Recall} = \frac{\text{Relevant documents} \cap \text{Retrieved documents}}{\text{Relevant documents}}$$

- When the recall measure is used, there is an assumption that all the relevant documents for a given query are known.
- Such an assumption is clearly problematic in a web search environment, but with smaller test collection of documents, this measure can be useful.
- It is not suitable for large volumes of log data.

- **Emphasis on users and their information needs**
- The users of a search engine are the ultimate judges of quality.
- This has led to numerous studies on how people interact with search engines and in particular, to the development of techniques to help people express their information needs.

- Text queries are often poor descriptions of what the user actually wants compared to the request to a database system, such as for the balance of a bank account.

- The figure summarizes the major issues involved in search engine design



The User Task

The user of a retrieval system has to translate his information need into a query in the language provided by the system.

With an information retrieval system, this normally implies specifying a set of words which convey the semantics of the information need.

With a data retrieval system, a query expression (such as, for instance, a regular expression) is used to convey the constraints that must be satisfied by objects in the answer set.

In both cases, we say that the user searches for useful information executing a *retrieval* task.

Information versus Data Retrieval

Data retrieval, in the context of an IR system, consists mainly of determining which documents of a collection contain the keywords in the user query which, most frequently, is not enough to satisfy the user information need.

In fact, the user of an IR system is concerned more with retrieving *information* about a subject than with retrieving data which satisfies a given query.

A data retrieval language aims at retrieving all objects which satisfy clearly defined conditions such as those in a regular expression or in a relational algebra expression.

For an information retrieval system, however, the retrieved objects might be inaccurate and small errors are likely to go unnoticed.

The main reason for this difference is that information retrieval usually deals with natural language text which is not always well structured and could be semantically ambiguous.

Information versus Data Retrieval

Data retrieval, while providing a solution to the user of a database system, does not solve the problem of retrieving information about a subject or topic.

To be effective in its attempt to satisfy the user information need, the IR system must some how `interpret' the contents of the information items (documents) in a collection and rank them according to a degree of relevance to the user query.

This `interpretation' of a document content involves extracting syntactic and semantic information from the document text and using this information to match the user information need. The difficulty is not only knowing how to extract this information but also knowing how to use it to decide relevance. Thus, the notion of *relevance* is at the center of information retrieval. In fact, the primary goal of an IR system is to retrieve all the documents which are relevant to a user query while retrieving as few non-relevant documents as possible.

Information Retrieval	Data Retrieval
The software the program that deals with the organization, storage, retrieval, and evaluation of information from document repositories particularly textual information.	Data retrieval deals with obtaining data from a database management system such as ODBMS. It is A process of identifying and retrieving the data from the database, based on the query provided by user or application.
Retrieves information about a subject.	Determines the keywords in the user query and retrieves the data.
Small errors are likely to go unnoticed.	A single error object means total failure.
Not always well structured and is semantically ambiguous.	Has a well-defined structure and semantics.
Does not provide a solution to the user of the database system.	Provides solutions to the user of the database system.
The results obtained are approximate matches.	The results obtained are exact matches.
Results are ordered by relevance.	Results are unordered by relevance.
It is a probabilistic model.	It is a deterministic model.

Information retrieval (*IR*) is the activity of obtaining *information system* resources that are relevant to an *information* need from a collection of those resources.

Searches can be based on full-text or other content-based indexing.

The **IR system** assists the users in finding the information they require but it does not explicitly return the answers to the question.

It notifies regarding the existence and location of documents that might consist of the required information.

Information retrieval also extends support to users in browsing or filtering document collection or processing a set of retrieved documents.

The system searches over billions of documents stored on millions of computers.

A spam filter, manual or automatic means are provided by Email program for classifying the mails so that it can be placed directly into particular folders.

An IR system has the ability to represent, store, organize, and access information items. A set of **keywords** are required to search. Keywords are what people are searching for in search engines.

The **IR system** assists the users in finding the information they require but it does not explicitly return the answers to the question.

It notifies regarding the existence and location of documents that might consist of the required information.

Information retrieval also extends support to users in browsing or filtering document collection or processing a set of retrieved documents.

The system searches over billions of documents stored on millions of computers.

A spam filter, manual or automatic means are provided by Email program for classifying the mails so that it can be placed directly into particular folders.

An IR system has the ability to represent, store, organize, and access information items. A set of **keywords** are required to search. Keywords are what people are searching for in search engines.

To Describe IR system, we use a simple and generic software architecture.

The First step in setting up an IR system is to assemble the document collection which can be provide or crawled from the web.

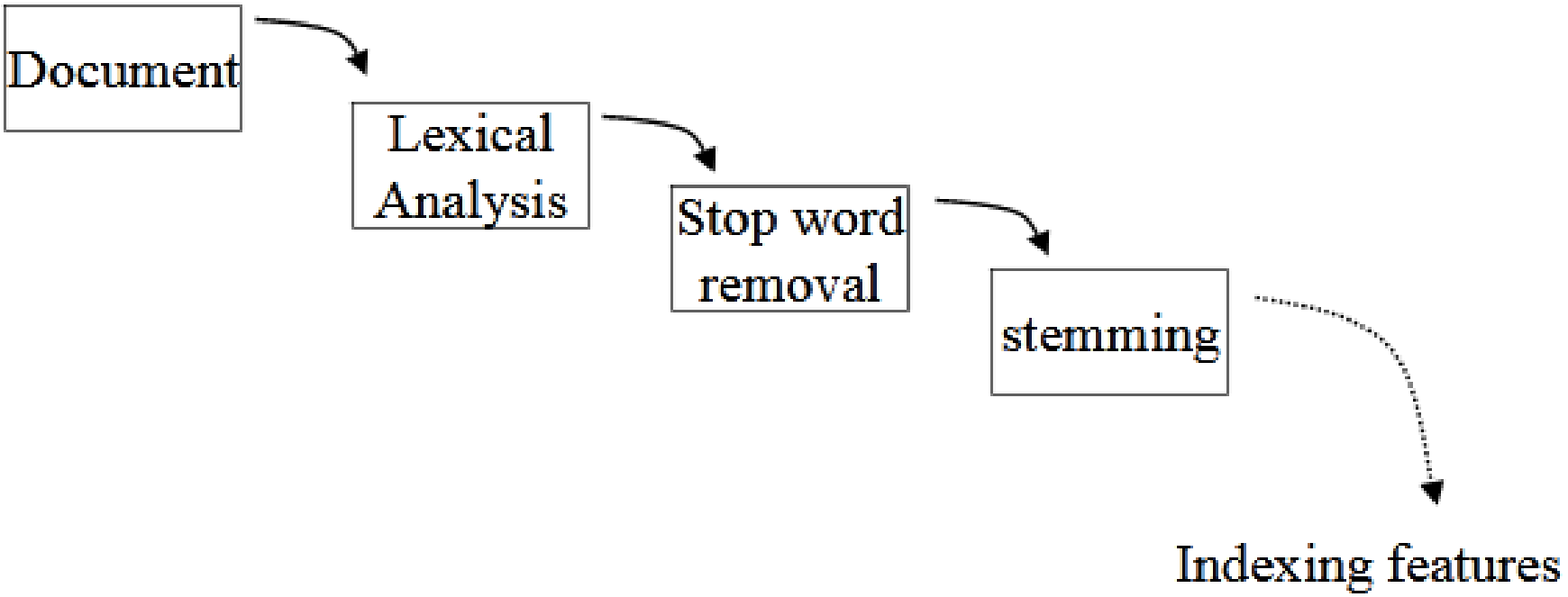
In the second case a crawler module is responsible for collecting the documents

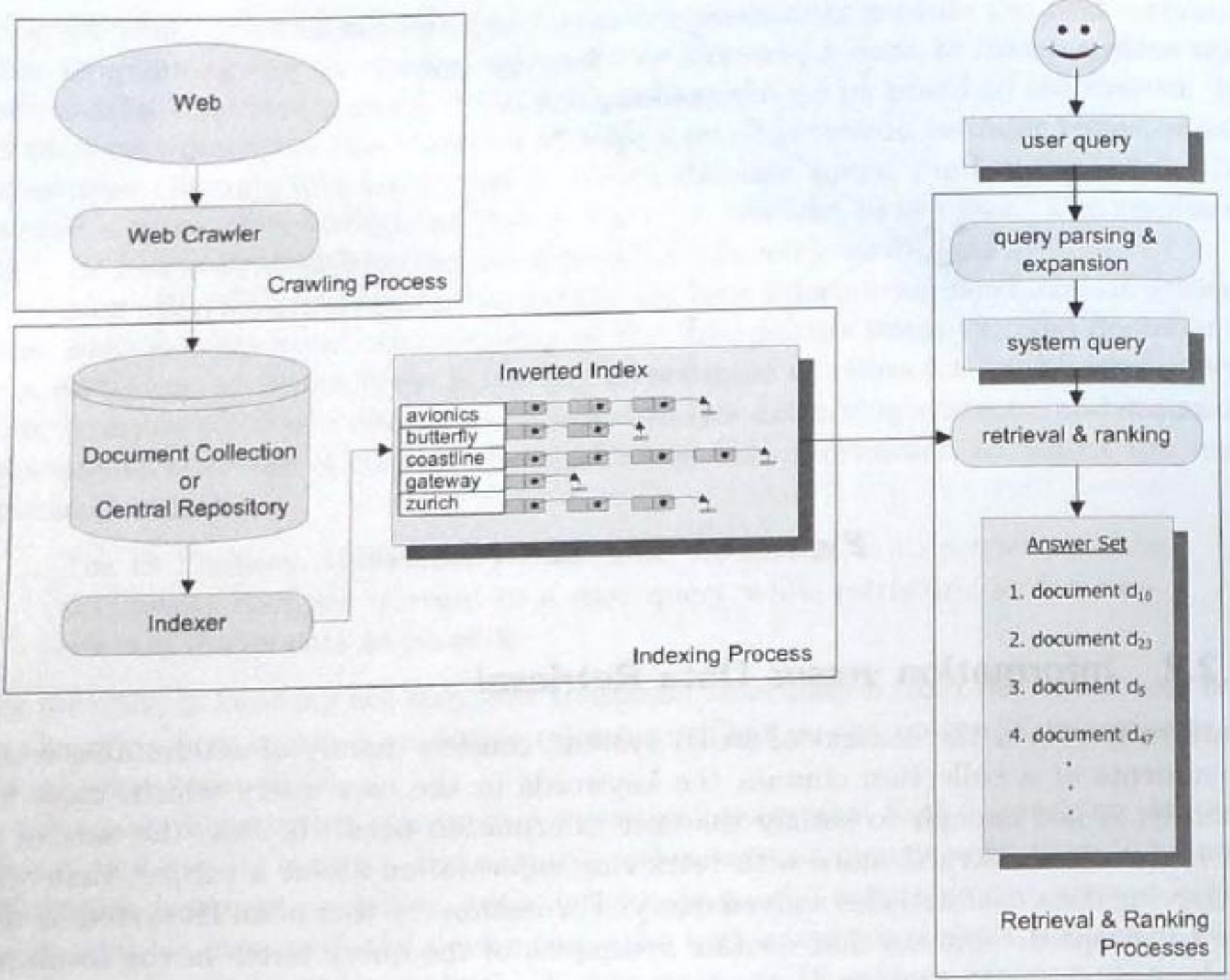
The Document collection is stored in the disk usually referred to as the central repository.

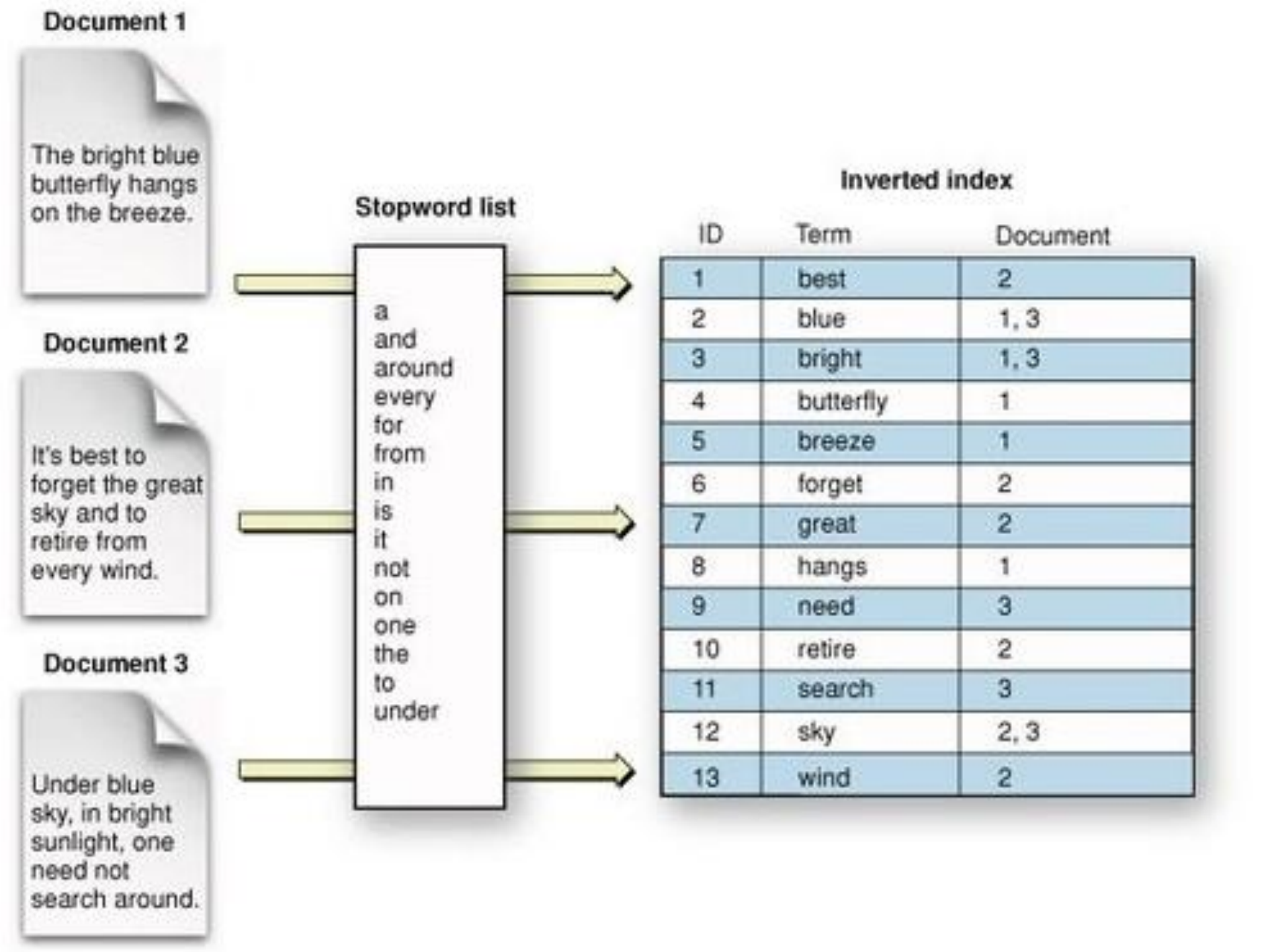
The document in the central repository need to be indexed for fast retrieval and ranking.

The most used index structure is an inverted index composed of all the distinct words of the collection and for each word , a list of the documents that contain it.

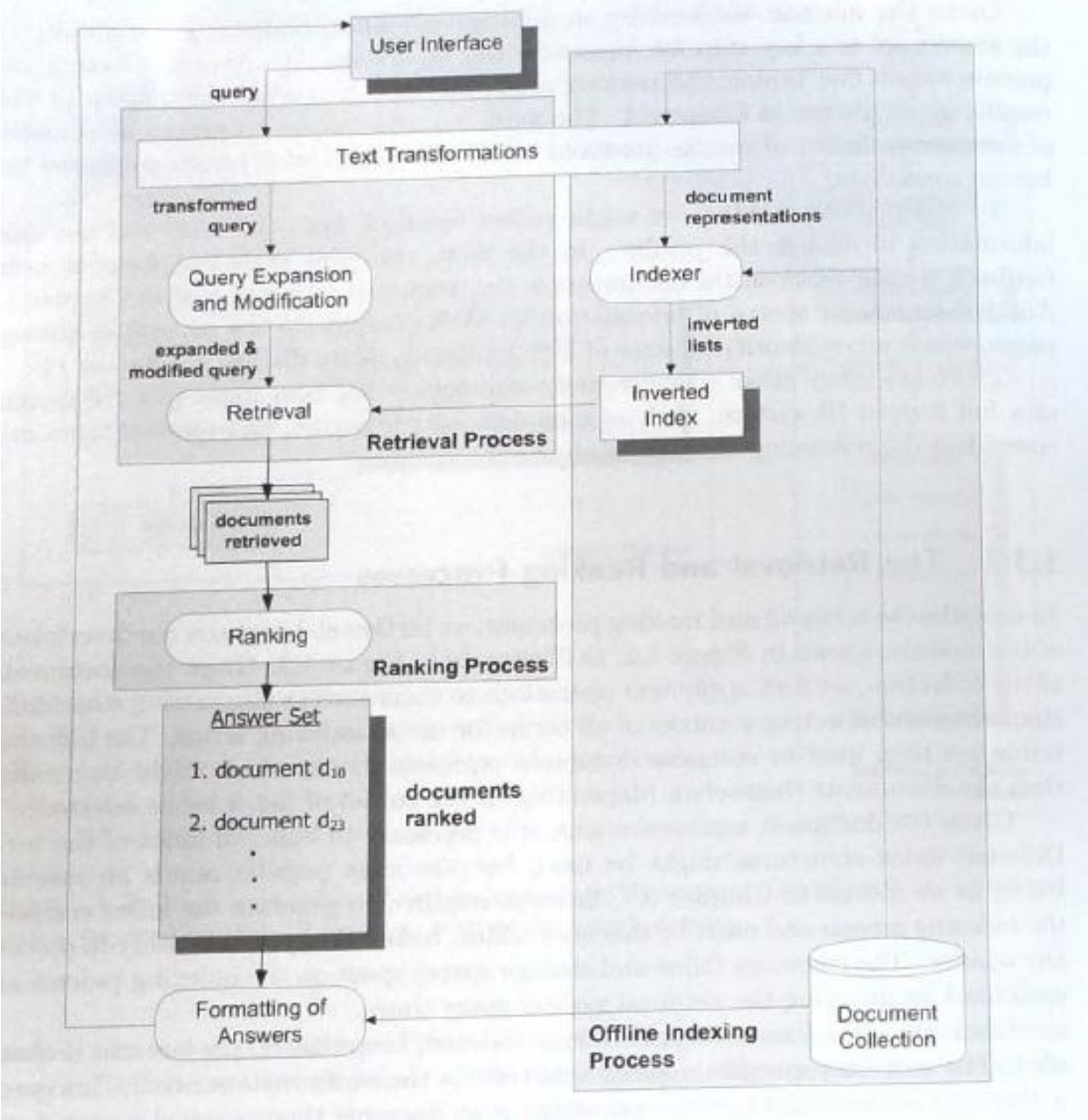
Process View




Web
Web Crawler
Crawling Process
Document Collection
or
Central Repository
Indexer
Inverted Index
avionics
butterfly
coastline
gateway
zurich
Indexing Process
user query
query parsing &
expansion
system query
retrieval & ranking
Answer Set
1. document d_{10}
2. document d_{23}
3. document d_5
4. document d_{47}
Retrieval & Ranking
Processes


Document 1
The bright blue butterfly hangs on the breeze.
Document 2
It's best to forget the great sky and to retire from every wind.
Document 3
Under blue sky, in bright sunlight, one need not search around.
Stopword list
a
and
around
every
for
from
in
is
it
not
on
one
the
to
under
Inverted index
ID
Term
Document
1
best
2
2
blue
1, 3
3
bright
1, 3
4
butterfly
1
5
breeze
1
6
forget
2
7
great
2
8
hangs
1
9
need
3
10
retire
2
11
search
3
12
sky
2, 3
13
wind
2

- A majority of search engines use **ranking** algorithms to provide users with accurate and relevant results.
-



The Web

The e-Publishing era

- Since its inception the Web became a huge success . The number of web pages now far exceeds 20 billion and the number of web users in the world exceeds 1.7 billion

- In the world of the Web, this is no longer the case
- People can now publish their ideas on the web reach millions of people over night , without paying anything for it and without having to convince the editorial board of a large publishing company.

How the Web Changed Search

- Web search is today the most prominent application of IR and its techniques—the ranking and indexing components of any search engine are fundamentally IR pieces of technology.

- The first major impact of the Web on search is related to the characteristics of the document collection itself

- The Web is composed of pages distributed over millions of sites and connected through hyperlinks
- This requires collecting all documents and storing copies of them in a central repository, prior to indexing
- This new phase in the IR process, introduced by the Web, is called crawling

- The second major impact of the Web on search is related to:

- The size of the collection
- The volume of user queries submitted on a daily basis
- As a consequence, performance and scalability have become critical characteristics of the IR system

- The third major impact:
- In a very large collection, predicting relevance is much harder than before
- Fortunately, the Web also includes new sources of evidence
- Ex: hyperlinks and user clicks in documents in the answer set

- The fourth major impact derives from the fact that the Web is also a medium to do business
- Search problem has been extended beyond the seeking of text information to also encompass other user needs
- Ex: the price of a book, the phone number of a hotel, the link for downloading a software

- The fifth major impact of the Web on search is Web spam

- Web spam: abusive availability of commercial information disguised in the form of informational content
- This difficulty is so large that today we talk of Adversarial Web Retrieval

- Web spam is annoying to search engine users and disruptive to search engines; therefore, most commercial search engines try to combat web spam.

Practical Issues in the web

Security

- Electronic commerce is a major trend on the web nowadays and one which has benefited million of peoples .
- In an Electronic transaction the buyer usually submits to the vendor credit information to be used for charging purposes .

Privacy

- Frequently people are willing to exchange information as long as it does not become public.
- The Reasons are many but most common one is to protect oneself misuse of private information by third parties

Added Protection

Please submit your One Time Password.

Merchant:	AMAZON PAY INDIA PRIVATE
Amount:	**INR 569.00**
Date:	24/12/2020
Card number:	4573XXXXXXXX1006
OTP:	

Regenerate OTP

*** Your One Time Password (OTP) has been sent to your mobile number ending XXX 1548**

Submit

Help Exit

Practical Issues in the web

Copyright and Patent rights

- It is far from clear how the wide spread on the web affects copyright and patents
- This is Important because it affects the business of building up and deploying large digital libraries .

- For instance, is a site which supervises all the information it posts acting as a publisher?
- And if so, is it responsible for a misuse of the information it posts (even if it is not the source)?

Practical Issues in the web

Interest include scanning, optical character recognition (OCR), and cross-language retrieval (in which the query is in one language but the documents retrieved are in another language).

How to do Search

- In our modern information society, data, facts, and knowledge have a much higher priority than they were half a century ago.

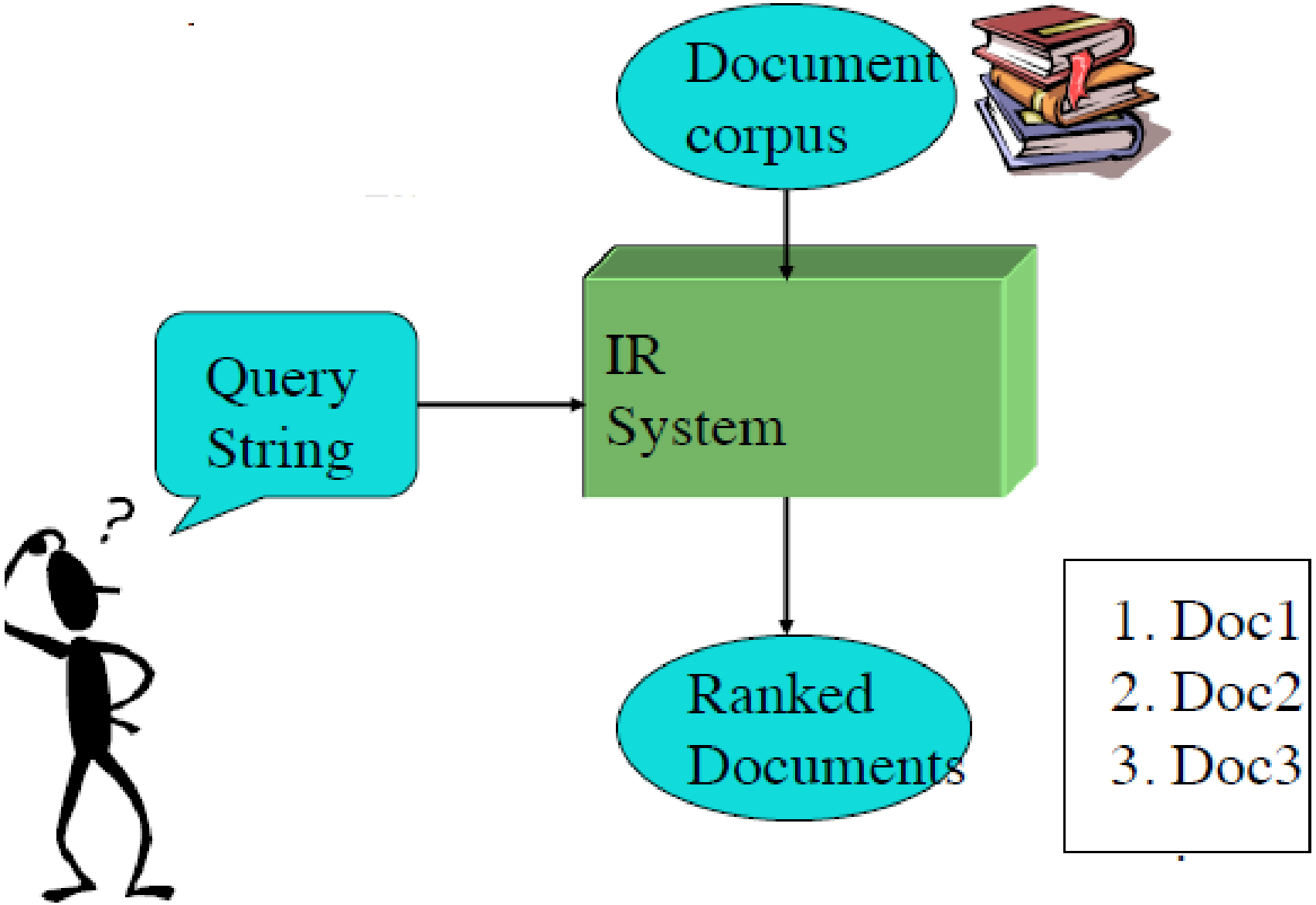
- Thanks to the internet, information is more and more accessible. When we access information, it has been retrieved from online sources, which is where search engines come in.

- How do they find the information they provide us? The answer is called “Information Retrieval”.

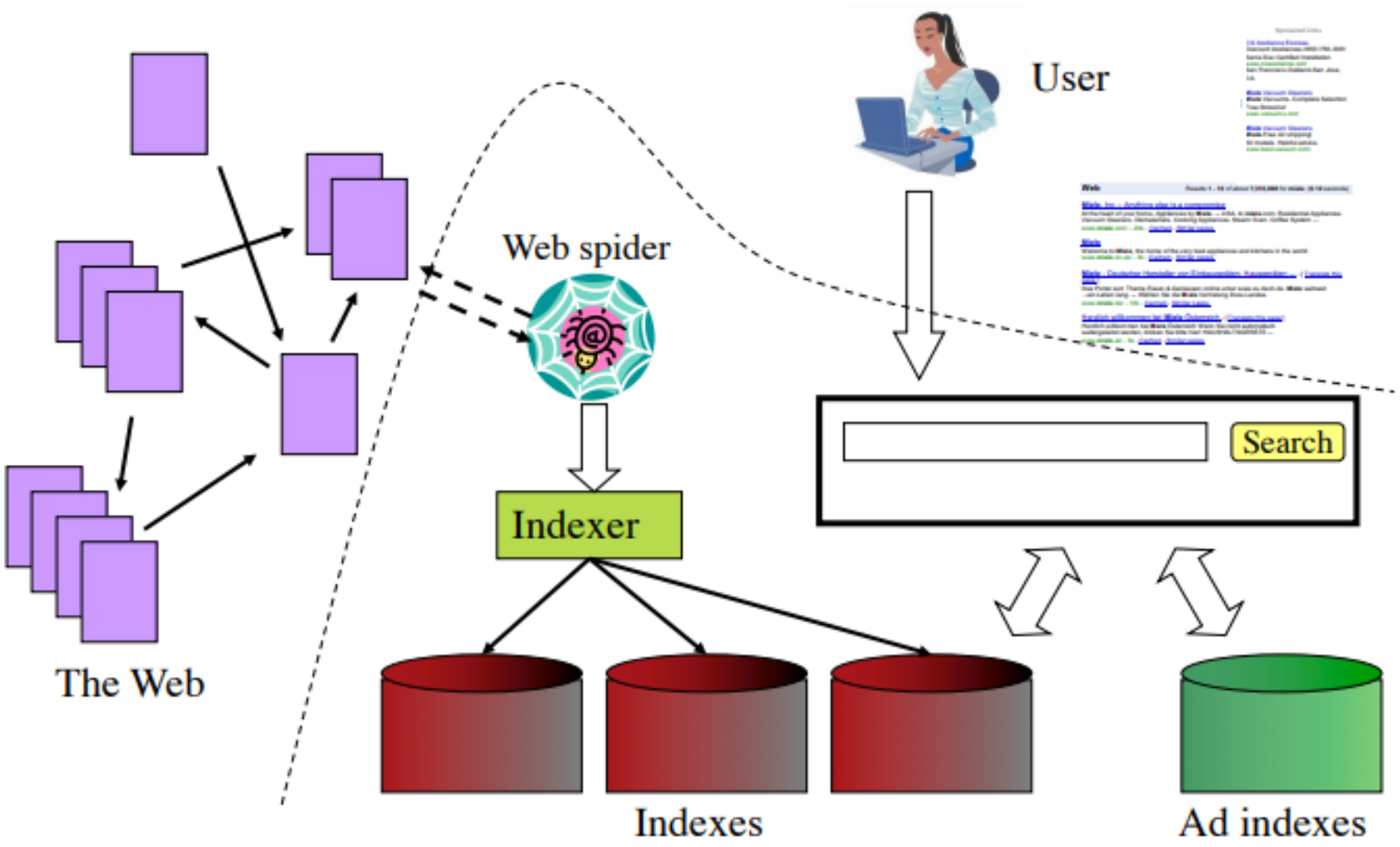
- Gathering information – more precisely, **information recovery** – is a discipline in computer science and information science and, above all, of great importance for search engines.

- Using complex information retrieval systems, they recognize the intentions that are behind specific search terms and find relevant data on search queries.

How to do Search

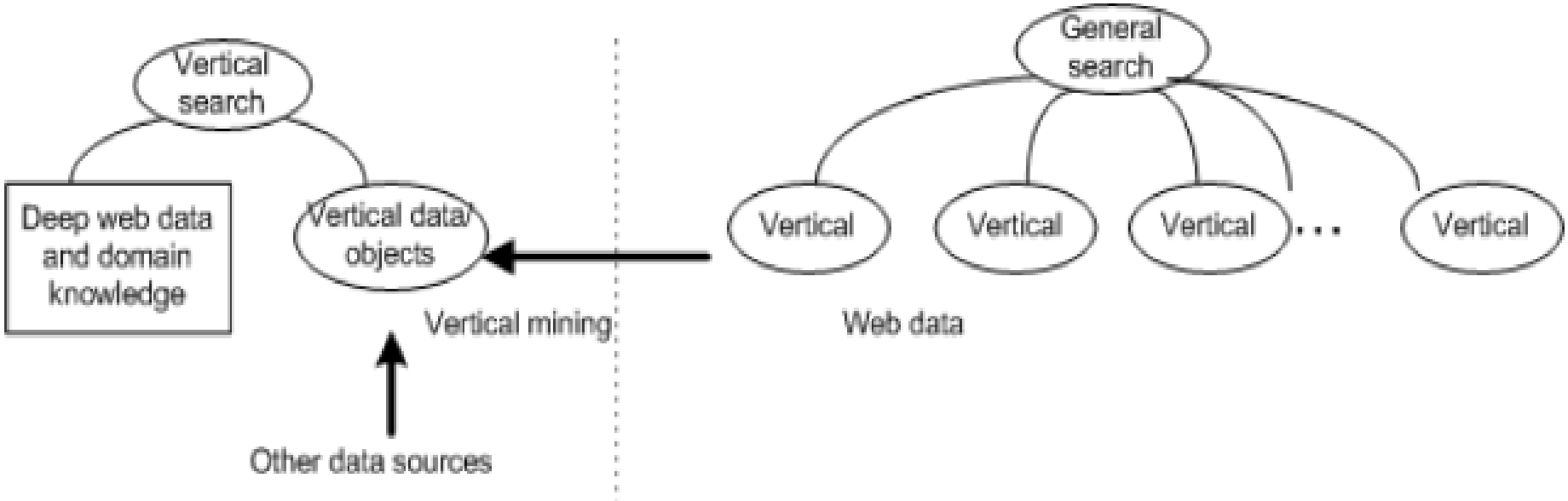


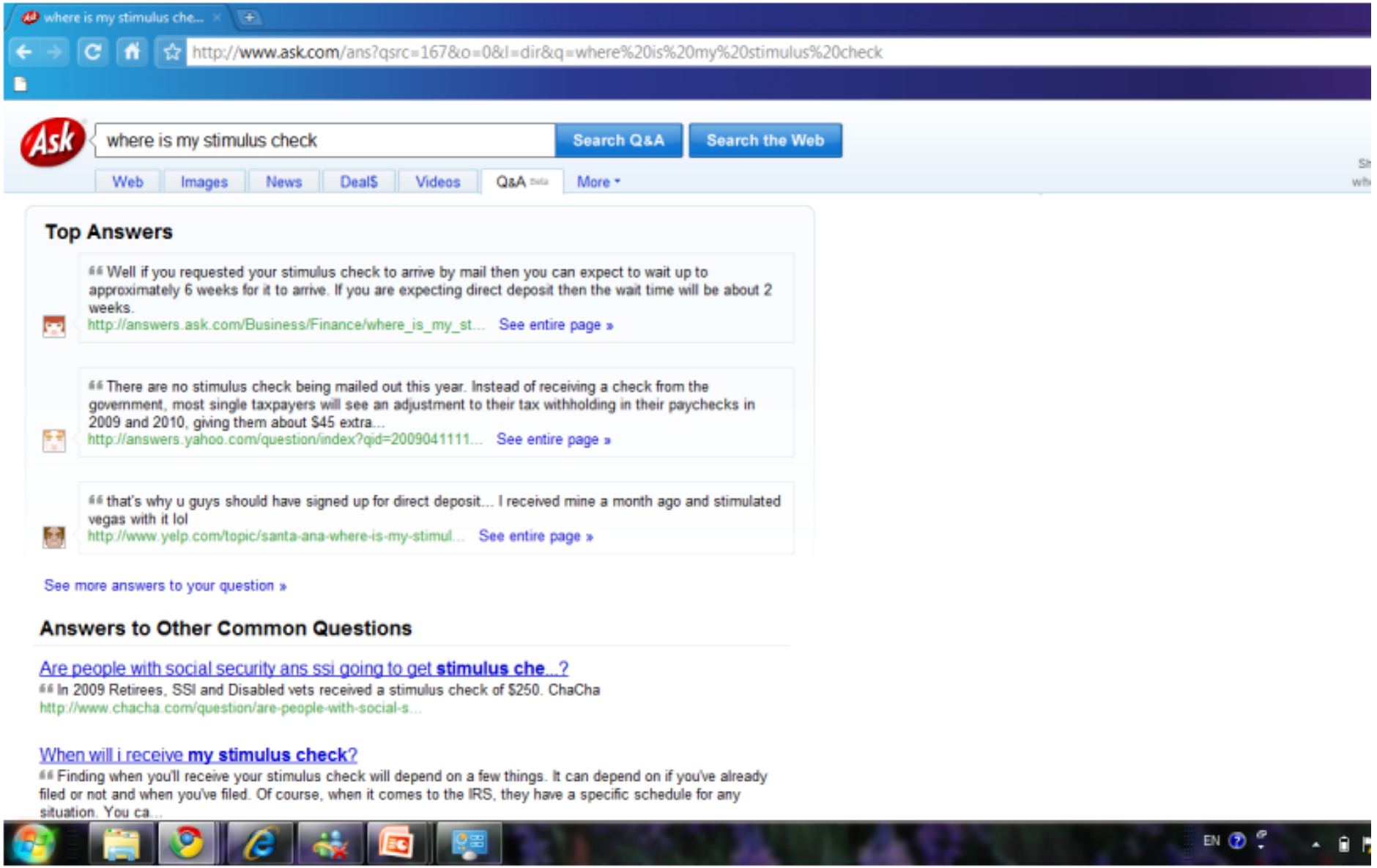
How to do Search



General Search vs. Vertical Search

- General Search: identify relevant information with a horizontal/exhaustive view of the world.
- Vertical Search:
- A vertical search refers to search on a specific topic area or a specific segment of an overall search. There are vertical or specialized search engines. An example of a vertical search is Google Image Search. Typical vertical search queries include shopping, travel, cars, medical information, and books.




where is my stimulus che...
http://www.ask.com/ans?qsrc=167&o=0&l=dir&q=where%20is%20my%20stimulus%20check
Ask
where is my stimulus check
Search Q&A
Search the Web
Web
Images
News
Deal$
Videos
Q&A beta
More
Top Answers
Well if you requested your stimulus check to arrive by mail then you can expect to wait up to approximately 6 weeks for it to arrive. If you are expecting direct deposit then the wait time will be about 2 weeks.
http://answers.ask.com/Business/Finance/where_is_my_st... See entire page »
There are no stimulus check being mailed out this year. Instead of receiving a check from the government, most single taxpayers will see an adjustment to their tax withholding in their paychecks in 2009 and 2010, giving them about $45 extra...
http://answers.yahoo.com/question/index?qid=2009041111... See entire page »
that's why u guys should have signed up for direct deposit.... I received mine a month ago and stimulated vegas with it lol
http://www.yelp.com/topic/santa-ana-where-is-my-stimul... See entire page »
See more answers to your question »
Answers to Other Common Questions
Are people with social security ans ssi going to get stimulus che...?
In 2009 Retirees, SSI and Disabled vets received a stimulus check of $250. ChaCha
http://www.chacha.com/question/are-people-with-social-s...
When will i receive my stimulus check?
Finding when you'll receive your stimulus check will depend on a few things. It can depend on if you've already filed or not and when you've filed. Of course, when it comes to the IRS, they have a specific schedule for any situation. You ca...
EN

Horizontal Searching

- Horizontal searching can be thought of as a “general” search.
- This type of searching goes through the whole web covering a wide range of topics and media subjects.
- A horizontal search will often produce a large number of results.
- The ranking for results varies based on the search engines algorithm but the results try to satisfy as many search queries as possible.
- An example of a horizontal search would be looking for “Toronto SEO” in Google.
- You’ll get tons of results for web pages that have both SEO and Toronto on them.

Search Interfaces Today

- At the heart of the typical search session is a cycle of query specification , inspection of retrieval results , and query reformulation.
- How does an information seeking session begin in online information systems?

- For a user of online information systems the most common way to start a search session is to access a web browser and use a web search engine

- Another method is to select a Web site from a personal collection of already-visited sites.
  - which are typically stored in a browser's bookmark

- Online bookmark systems are popular among a smaller segment of users
- Ex: Delicious.com
- Web directories are also used as a common starting point, but have been largely replaced by search engines

Query Specification

- Once a search starting point has been selected , the primary methods for a searcher to express their information need are either entering words into a search entry from or selecting links from a directory or other information organization display .

- In the query specification part of the information access process, the searcher expresses an information need by converting their internalized, abstract concepts into language, and then converting that expression of language into a query format that the search system can make use of.

- Web ranking has gone through three major phases

- **Statistical Ranking**

- **Conjunctive Queries**

- **Sophisticated searcher found that longer Queries**

The two main dimensions for the query specification process are:

[1] The kind of information the searcher supplies.

Query specification input spans a spectrum from full natural language sentences, to keywords and key phrases, to syntax-heavy command language-based queries.

[2] The interface mechanism the user interacts with to supply this information.

These include command line interfaces, graphical entry form-based interfaces, and interfaces for navigating links.

- Typically in web queries today the text is very short, consisting of one two three words .
- Multiword queries are often meant to be construed as a phase but can also consist of multiple topics.
- Short queries reflect the standard usage scenario in which the user “tests the waters” to see what the search engine returns in response to their short query.
- If the results do not look relevant then the user reformulates their query.

Query Specification interfaces

- The standard interface for a textual query is a search box entry form , in which the user types a query activated by hitting the return key on the keyboard or selecting a button associates with the form.
- To look for a book on a topic in an online library catalog, the searcher was restricted to the text in the title or the few subject labels that the librarian who had catalogued the book had used to describe it
- Some entry forms are divided into multiple components allowing for a more general free text query followed by a form that filters the query in some way.
- With the rise of the Web came the dominance of keywords as the primary query input type.
- Keyword queries consist of a list of one or more words or phrases -- rather than full natural language statements -- whose intention is to find documents containing those words that are likely to be relevant to the user's information need.

Query Specification interfaces

- Example
  - English keyword queries (from Google Trends) include flip cam, fresh chilli paste recipes, and video game addiction.
  - Some keyword queries consist of lists of different words and phrases, which together suggest a topic (e.g., early voting Florida).

Query Specification interfaces

- Dynamic Query Suggestion
  - In dynamic query suggestion interfaces , the display of the matches varies .
  - In most cases the user must move the mouse down to the desired suggestion in order to select it .

  - Suggestions can be derived from several sources
  - In some cases the list is taken from the user's own query history
  - It is based on popular query issued by the users
  - The list can be derived from a set of metadata that web site's.

  - Another form of query specification consists of choosing from a display of information, typically in the form of hyperlinks or saved bookmarks .

Retrieval Results and Display

- In information retrieval systems and digital libraries, result presentation is a very important aspect.

- When displaying search results either the documents must be shown in full or else the searcher must be presented with some kind of representation of the content of these documents

- Demonstrate that only a ranked list of documents, thought commonly used by many retrieval systems and digital libraries, is not the best way of presenting retrieval results.

Retrieval Results and Display

- We believe, in many situations, an estimated relevance probability score or an estimated relevance score should be provided for every retrieved document by the information retrieval system/digital library.

- With such information, the usability of the retrieval result can be improved, and the Euclidean distance can be used as a very good system-oriented measure for the effectiveness of retrieval results.

Query Reformulation

- After a Query is specified and results have been produced a number of tools exist to help the user reformulate their query, or take their information seeking process in new direction .

Retrieval Results Display

- When displaying search results either the documents must be shown in full or else the searcher must be presented with some kind of representation of the content of those documents .

- In web search the page title is usually shown prominently along with URL and sometimes other meta data.

- The text summary (also called the abstract , extract , excerpt , and snippet) containing text extracted from the document is also critical for assessment

- Currently the standard results display is a vertical list of textual summaries and is sometimes referred to as SERP(Search Engine Result Page).

Retrieval Results Display

- Search engine results pages are web pages served to users when they search for something online using a search engine, such as Google.

- The user enters their search query (often using specific terms and phrases known as *keywords*), upon which the search engine presents them with a SERP.

- Every SERP is unique, even for search queries performed on the same search engine using the same keywords or search queries.

- This is because virtually all search engines customize the experience for their users by presenting results based on a wide range of factors beyond their search terms, such as the user's physical location, browsing history, and social settings.

Retrieval Results Display

- SERPs typically contain two types of content – “organic” results and paid results.

- Organic results are listings of web pages that appear as a result of the search engine’s algorithm .

- Search engine optimization professionals, commonly known as SEOs, specialize in optimizing web content and websites to rank more highly in organic search results.

Retrieval Results Display

There are three primary types of Internet search:

- Informational
  - Informational searches are those in which the user hopes to find information on a given topic, such as Abraham Lincoln.
  - It wouldn't make much sense to place ads or other types of paid results on a SERP like this, as the search query "Abraham Lincoln" has very low *commercial intent*;
  - the vast majority of searchers using this search query are not looking to buy something, and as such only informational results are displayed on the SERP.
- Navigational
  - Navigational queries are those in which the user hopes to locate a specific website through their search.
  - This may be the case for individuals searching for a specific website, trying to locate a website whose URL they can no longer remember, or another type of navigational objective.

Retrieval Results Display

There are three primary types of Internet search:

- Transactional
  - Finally, transactional searches are those in which paid results are most likely to be displayed on the SERP.
  - Transactional searches have high commercial intent, and search queries leading to transactional SERPs may include keywords such as “buy” and other terms that suggest a strong desire to make a purchase.

Retrieval Results Display

Paid Results

- In contrast to organic results, paid results are those that have been paid to be displayed by an advertiser.
- In the past, paid results were almost exclusively limited to small, text-based ads that were typically displayed above and to the right of the organic results.
- Today, however, paid results can take a wide range of forms, and there are dozens of advertising formats that cater to the needs of advertisers.

Retrieval Results Display

- In some cases summaries are excerpts drawn from the full text that contain the query terms.

- In other cases specialized kind of metadata are shown in additional to standard textual results using a technique known as blended results .

Query Reformulation

- **Query reformulation** is the process of iteratively modifying a **query** to improve the quality of a search engine results, in order to satisfy one's information need.

- aim of such process is to satisfy the user information need, usually improving the quality and recall of the results obtained using the original user query.

- This feature is explicitly supported by some search engines suggesting related queries or providing different completions of the initial user query.

- Query reformulation also reflects the interplay between the surface and deeper levels of user interaction.

Query Reformulation

- Moreover, other search engines also support query reformulation in an implicit manner, by expanding the original query with terms related to their keywords
- One of the most important query reformulation techniques consists of showing terms related to the query or to the documents retrieved response to the query .

Visualization in Search Interfaces

- The goal of **information visualization** is to translate abstract **information** into a visual form that provides new insight about that **information**. **Visualization** has been shown to be successful at providing insight about data for a wide range of tasks.

- This Concept primarily about search of textual information .

- Text as a representation is highly effective for conveying abstract information but reading and even scanning text is cognitively taxing activity and must be done in a linear fashion

Visualization in Search Interfaces

- By contrast images can be scanned quickly and the visual system perceives information in parallel.

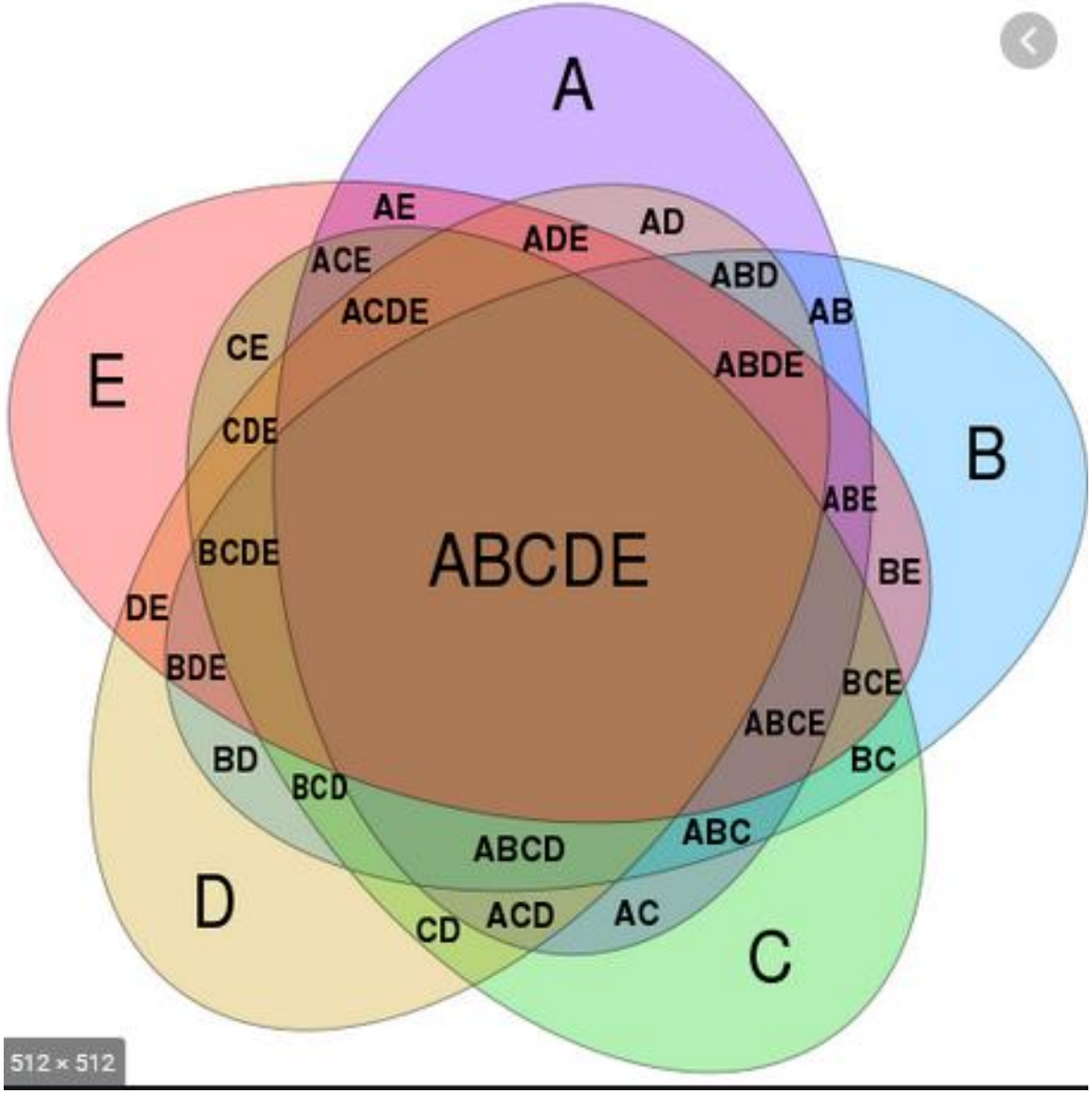
- People are highly attuned to images and visual information and pictures and graphics can be captivating and appealing

- Over the last few years , Information visualization has become a common presence in news reporting and financial analysis and creative innovative ideas about how to visualization information have blossomed and spread throughout the web

Visualization in Search Interfaces

Experimentation with visualization for search has been primarily applied in the following ways .

i. Visualizing Boolean syntax
ii. Visualizing Query Terms with Retrieval Results
iii. Visualizing relationship among words and Documents
iv. Visualizing for Text mining


A
B
C
D
E
AE
ADE
AD
ACE
ABD
AB
ACDE
CE
ABDE
CDE
ABE
BCDE
ABCDE
BE
DE
BDE
BCE
ABCE
BD
BC
BCD
ABC
ABCD
ACD
AC
CD
512 × 512

Visualizing Boolean syntax

Boolean query syntax is difficult for most users and is rarely used in web search

- The Venn diagram interface was significantly faster and produced significantly fewer errors than the textual query language

For many years researchers have experimented with how to visualize Boolean query specification in order to make it more understandable

Common approach is to show Venn diagrams visually.

Hertzum and frokjaer found that a simple Venn diagram representation produced more accurate results than Boolean syntax.

A more flexible version of this idea was seen in the VQuery system Each query term is represented by a circle or oval and a intersection among circles indicates ANDing (conjoining) of terms.

Vquery(graphical query interface) represented disjunction by sets of circles within an active area of a canvas and negation by deselecting a circle within the active area

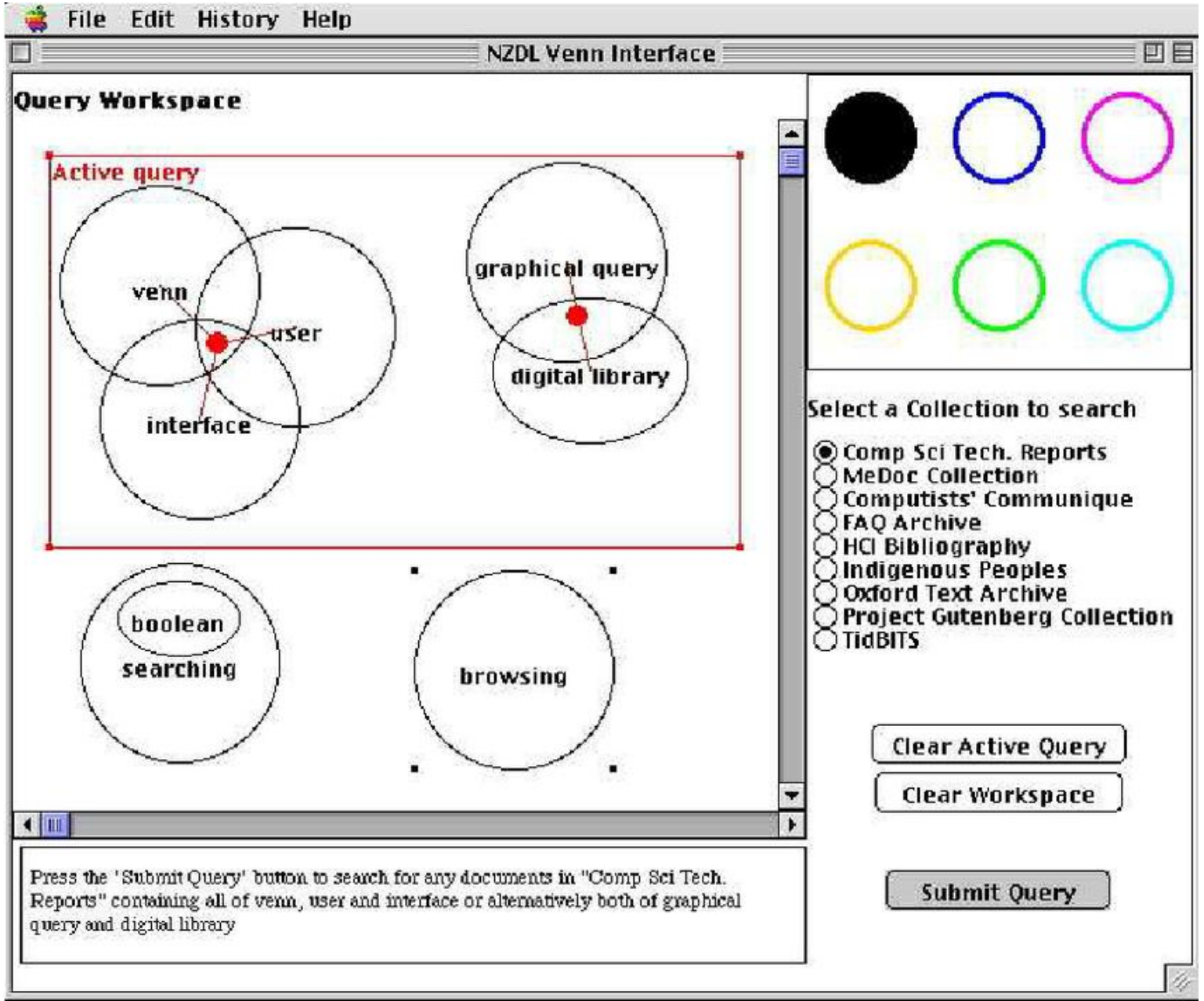
Visualizing Boolean syntax

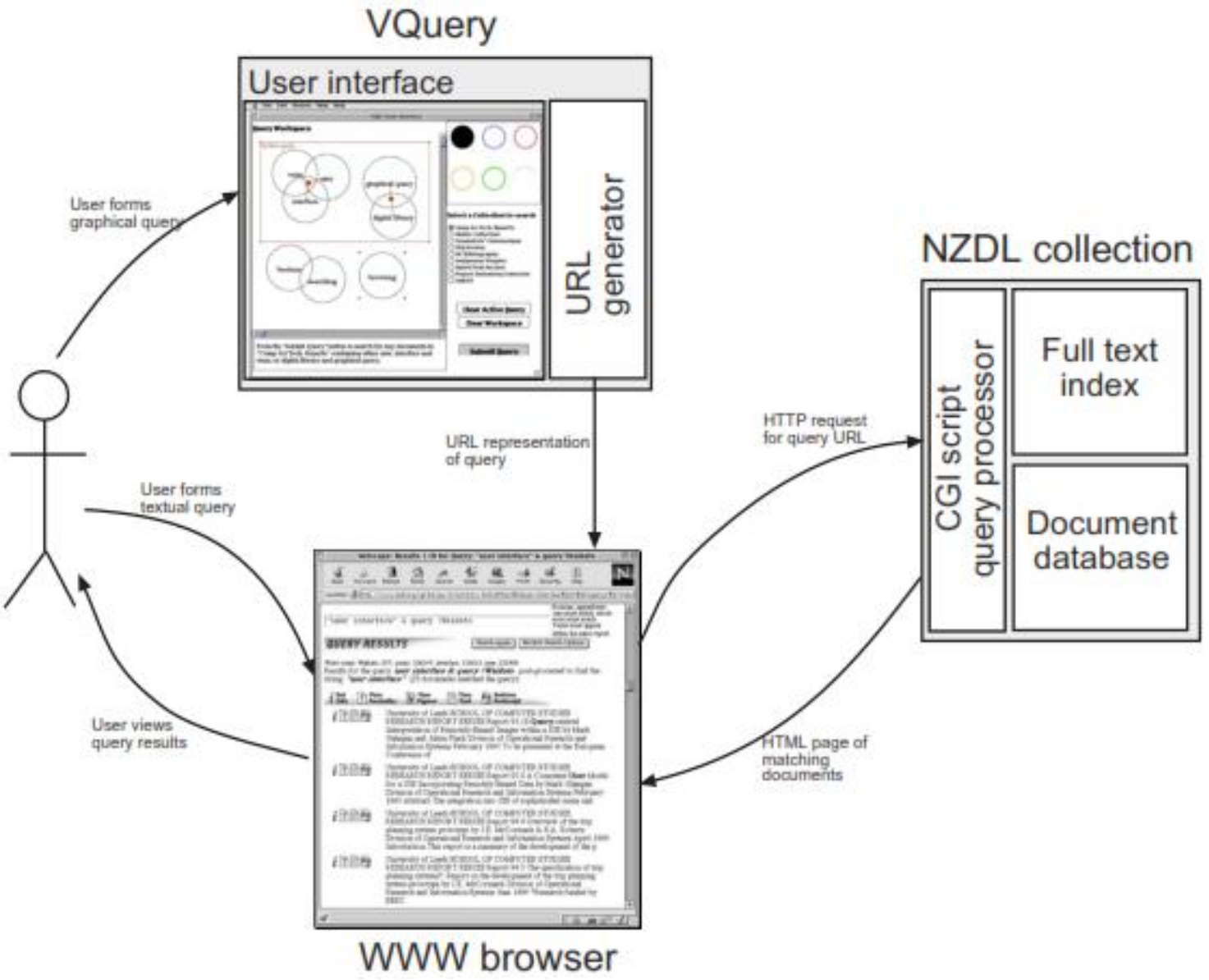
VQuery, is an alternative to the textual query interface

VQuery is implemented as a Java application which supports graphical creation and amendment of Boolean queries.

It does not deal with presentation of query results – that is left to the particular information source with which it interfaces.

Visualizing Boolean syntax

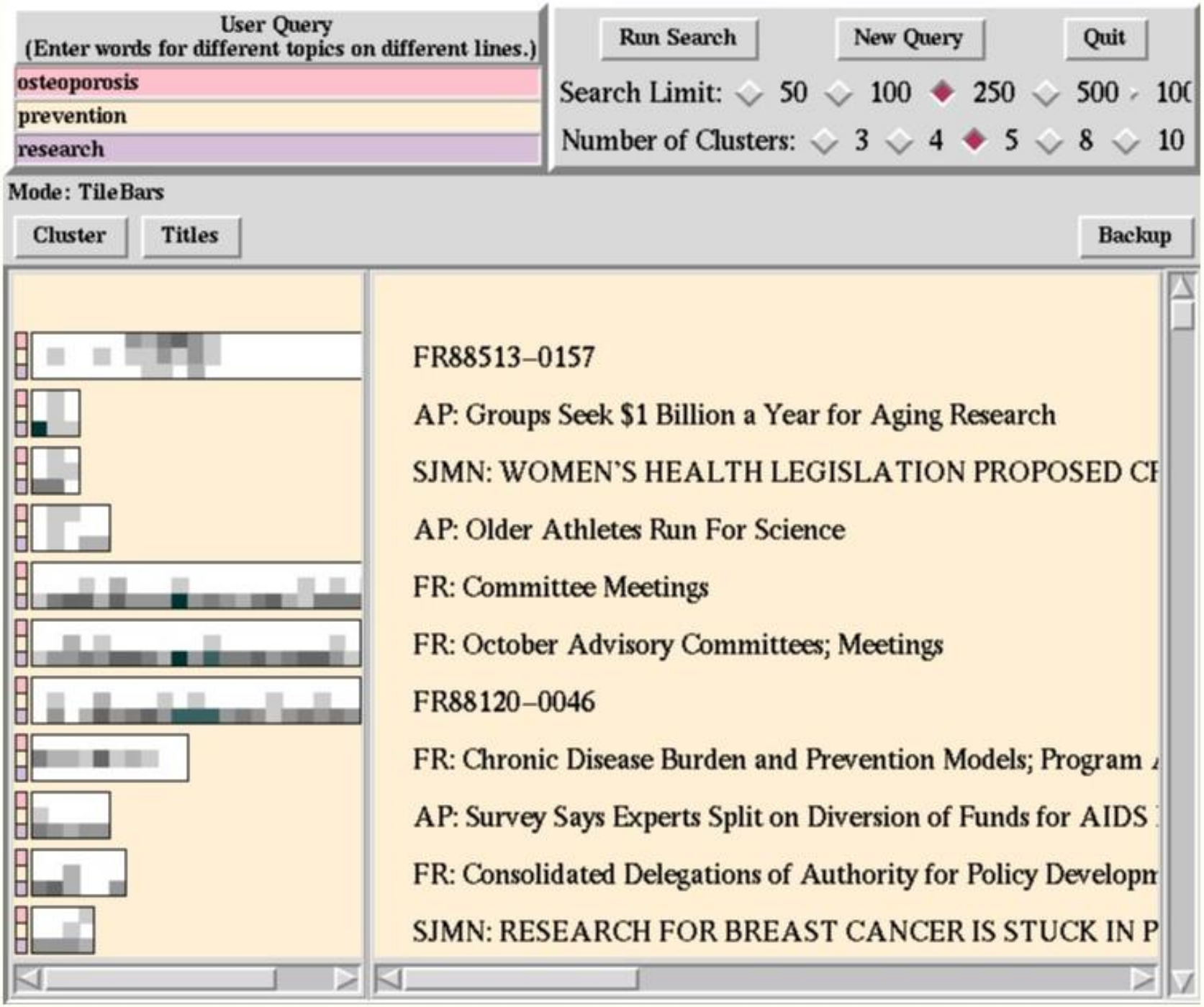



VQuery
User interface
URL generator
User forms graphical query
User forms textual query
User views query results
URL representation of query
HTTP request for query URL
HTML page of matching documents
NZDL collection
CGI script query processor
Full text index
Document database
WWW browser

Visualizing Query Terms with Retrieval Results

- Understanding the role played by the query terms within the retrieved documents can help with the assessment of relevance .
- In Standard search results listing , summary sentences are often selected that contain query terms , and the occurrence of these terms are highlighted or boldfaced where they appear in the title summary and URL.

- Exerimental visualizations have been designed that ake this relationship more explicit .
- One of the best known is the TileBars interface..
- A tileBar is a data visualization interface used in conjunction with a search feature, query terms, and expository text.


User Query
(Enter words for different topics on different lines.)
osteoporosis
prevention
research
Run Search
New Query
Quit
Search Limit: 50 100 250 500 100
Number of Clusters: 3 4 5 8 10
Mode: TileBars
Cluster
Titles
Backup
FR88513–0157
AP: Groups Seek $1 Billion a Year for Aging Research
SJMN: WOMEN'S HEALTH LEGISLATION PROPOSED CF
AP: Older Athletes Run For Science
FR: Committee Meetings
FR: October Advisory Committees; Meetings
FR88120–0046
FR: Chronic Disease Burden and Prevention Models; Program
AP: Survey Says Experts Split on Diversion of Funds for AIDS
FR: Consolidated Delegations of Authority for Policy Developm
SJMN: RESEARCH FOR BREAST CANCER IS STUCK IN P

Visualizing Query Terms with Retrieval Results

- The tileBar interface forces the user to input one or more queries,
- Generally of different topics and then displays each document's relevancy to the search through the use of rectangles,
- Which show the document's length through the size of the rectangle, the number of sections through the number of squares, and each section's frequency of each query through their color,
- Where a white square denotes a frequency of zero and a black, or the darkest color possible if not using black, square to denote a frequency of eight or more.
- Usually the search results are ranked by two metrics:
- Number of overlapping squares in each section and secondly by the overall frequency of terms in a given document.

Visualizing relationship among words and Documents

- Numerous visualization developers have proposed variations on the idea placing words and documents on a two dimension canvas
- Document analysis has recently been a hot research topic in the visualization field.
- Representing documents as individual objects and analyzing their content and relationships are helpful for many scenarios
- A tool that can provide instinctive visualizations of important aspects of the document collection is definitely helpful, especially for large collections.

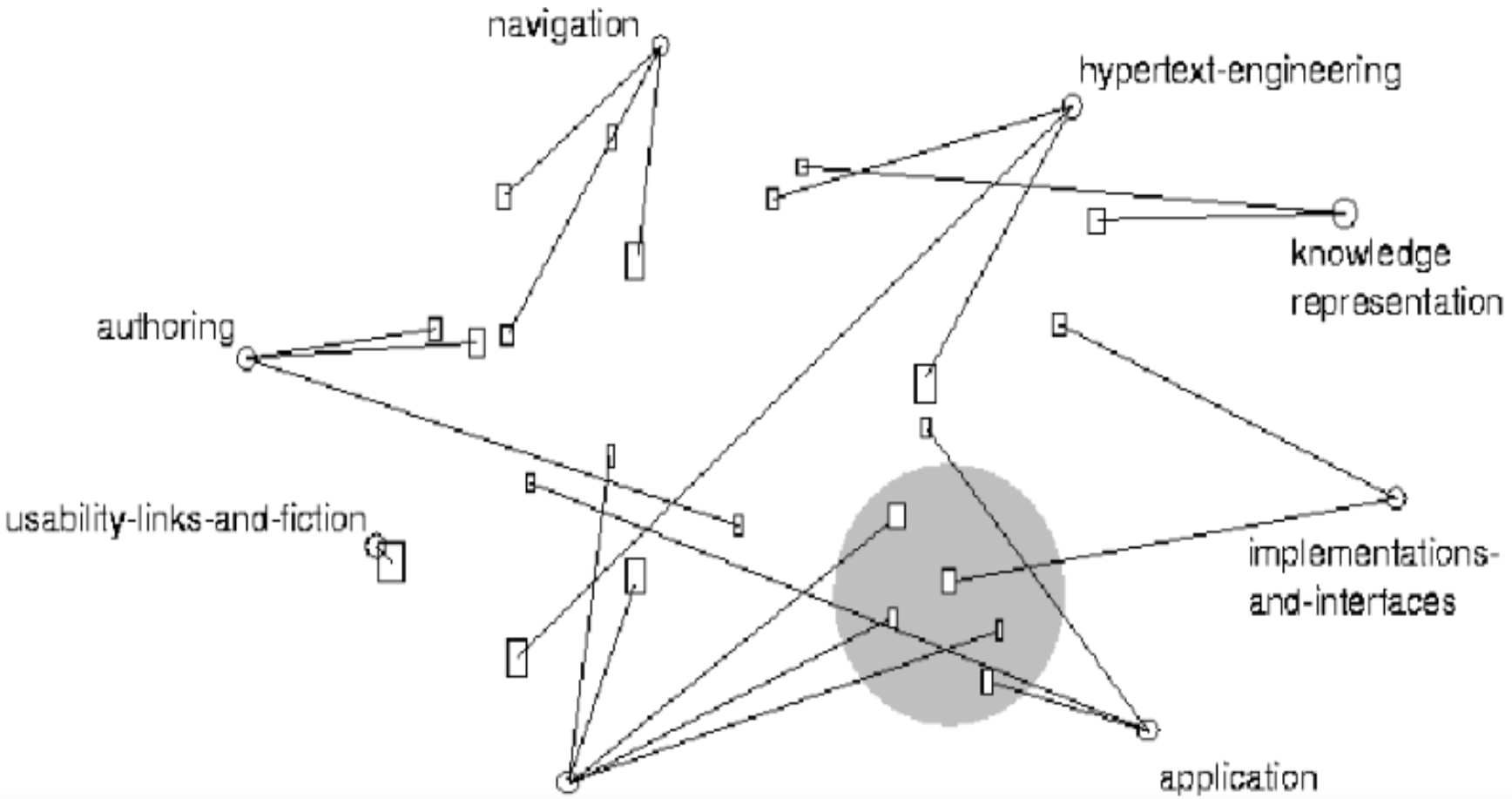


Figure 2.19: The VIBE display, in which query terms are laid out in a 2D space, and documents are arranged according to which subset of the text they share, from [1228].

Visualizing relationship among words and Documents

- This variation on clustering can be done to documents retrieved as a result of a query or documents that match a query can be highlighted within a proprocessed set of documents.

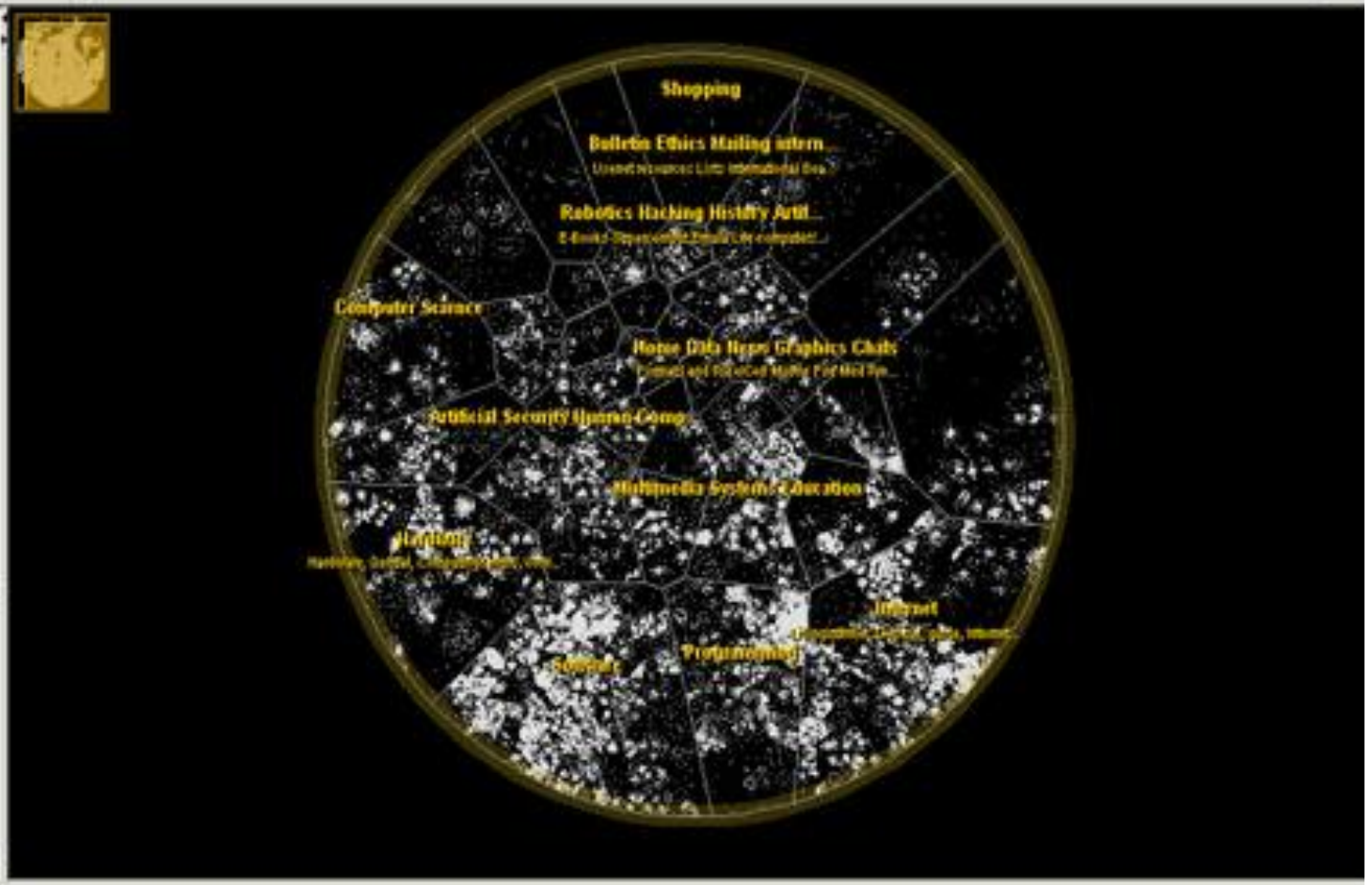
View_InfoSky_Video

InfoSky - Root

File

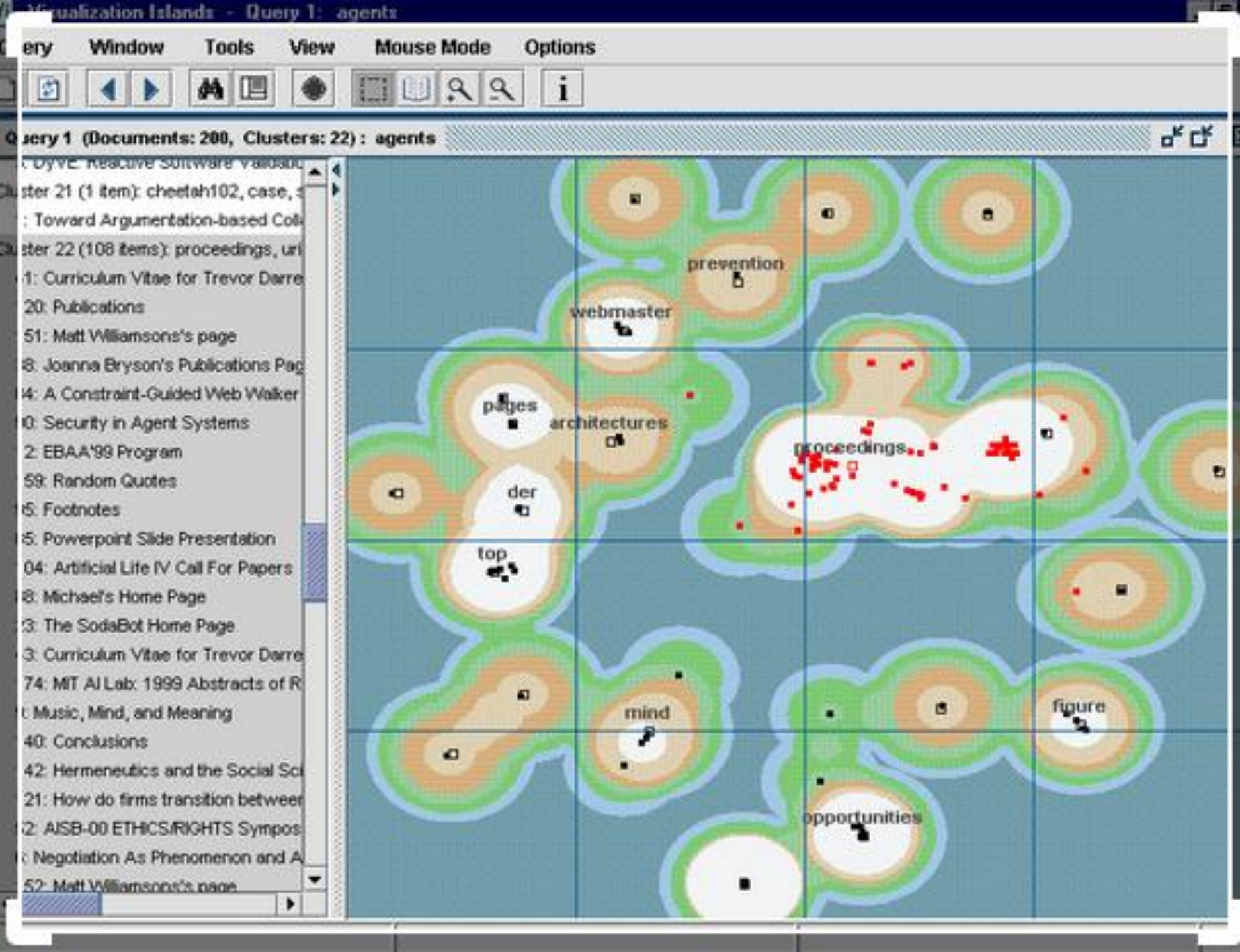
Root

- Root
  - Artificial_Intelligence
  - Home_Automation
  - Organizations
  - Robotics
  - Programming
  - Performance_and_Capacity
  - Multimedia
  - Software
  - Virtual_Reality
  - Data_Formats
  - Emulators
  - Bulletin_Board_Systems
  - Internet
  - Security
  - Hacking
  - Systems
  - Graphics
  - Data_Communications
  - Desktop_Publishing
  - Parallel_Computing
  - Algorithms
  - Consultants
  - Ethics
  - Supercomputing
  - Usenet
  - Computer_Science
  - Speech_Technology
  - E-Books



Details

Name	Size	Modified	Keywords
Artificial_Intelligence	1721 documents	Thu Jan 01 01:00:00 CET 1970	Artificial, Intelligence, artificial, intelligence, resources, links, Resources
Home_Automation	97 documents	Thu Jan 01 01:00:00 CET 1970	home, Home, automation, systems, networking, Internet, control
Organizations	308 documents	Thu Jan 01 01:00:00 CET 1970	Computing, technology, FOCUS, Circle, PenDragon, Humour, programming
Robotics	1105 documents	Thu Jan 01 01:00:00 CET 1970	robotics, Robotics, news, discussion, robot, site, robots
Programming	20409 documents	Thu Jan 01 01:00:00 CET 1970	
Performance_and_Capacity	78 documents	Thu Jan 01 01:00:00 CET 1970	performance, Performance, UNIX, test, Internet, System, computer
Multimedia	3476 documents	Thu Jan 01 01:00:00 CET 1970	multimedia, Multimedia, Authoring, design, content, resources, Multimediality


Visualization Islands - Query 1: agents
Window
Tools
View
Mouse Mode
Options
Query 1 (Documents: 200, Clusters: 22) : agents
Cluster 21 (1 item): cheetah102, case,
: Toward Argumentation-based Coll
Cluster 22 (108 items): proceedings, uri
1: Curriculum Vitae for Trevor Darre
20: Publications
51: Matt Williamsons's page
8: Joanna Bryson's Publications Pag
4: A Constraint-Guided Web Walker
0: Security in Agent Systems
2: EBAA'99 Program
59: Random Quotes
5: Footnotes
5: Powerpoint Slide Presentation
04: Artificial Life IV Call For Papers
8: Michael's Home Page
3: The SodaBot Home Page
3: Curriculum Vitae for Trevor Darre
74: MIT AI Lab: 1999 Abstracts of R
Music, Mind, and Meaning
40: Conclusions
42: Hermeneutics and the Social Sci
21: How do firms transition betwee
2: AISB-00 ETHICS/RIGHTS Sympos
Negotiation As Phenomenon and A
52: Matt Williamsons's page
prevention
webmaster
pages
architectures
proceedings
der
top
mind
figure
opportunities

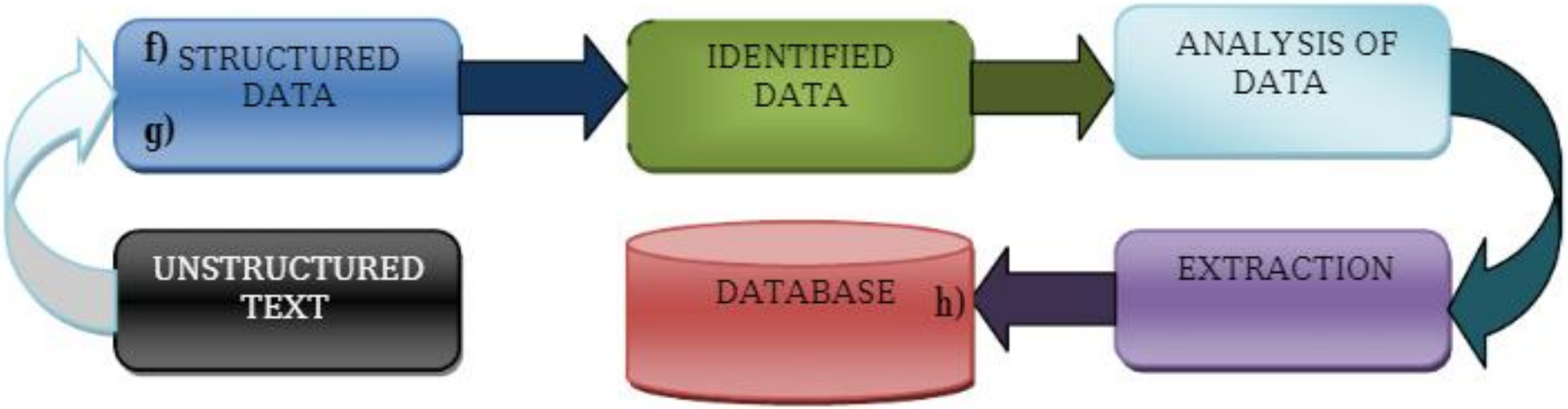
Visualization for Text Mining

- **Text mining** (also referred to as **text** analytics) is an artificial intelligence (AI) technology that uses natural language processing (NLP) to transform the free (unstructured) **text** in documents and databases into normalized, structured data suitable for analysis or to drive machine learning (ML) algorithms.

- *Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information from text.*

The five fundamental steps involved in text mining are:

- Gathering unstructured data from multiple data sources like plain text, web pages, pdf files, emails, and blogs, to name a few.
- Detect and remove anomalies from data by conducting pre-processing and cleansing operations. Data cleansing allows you to extract and retain the valuable information hidden within the data and to help identify the roots of specific words.
- For this, you get a number of text mining tools and text mining applications.
- Convert all the relevant information extracted from unstructured data into structured formats.
- Analyze the patterns within the data via the Management Information System (MIS).
- Store all the valuable information into a secure database to drive trend analysis and enhance the decision-making process of the organization.

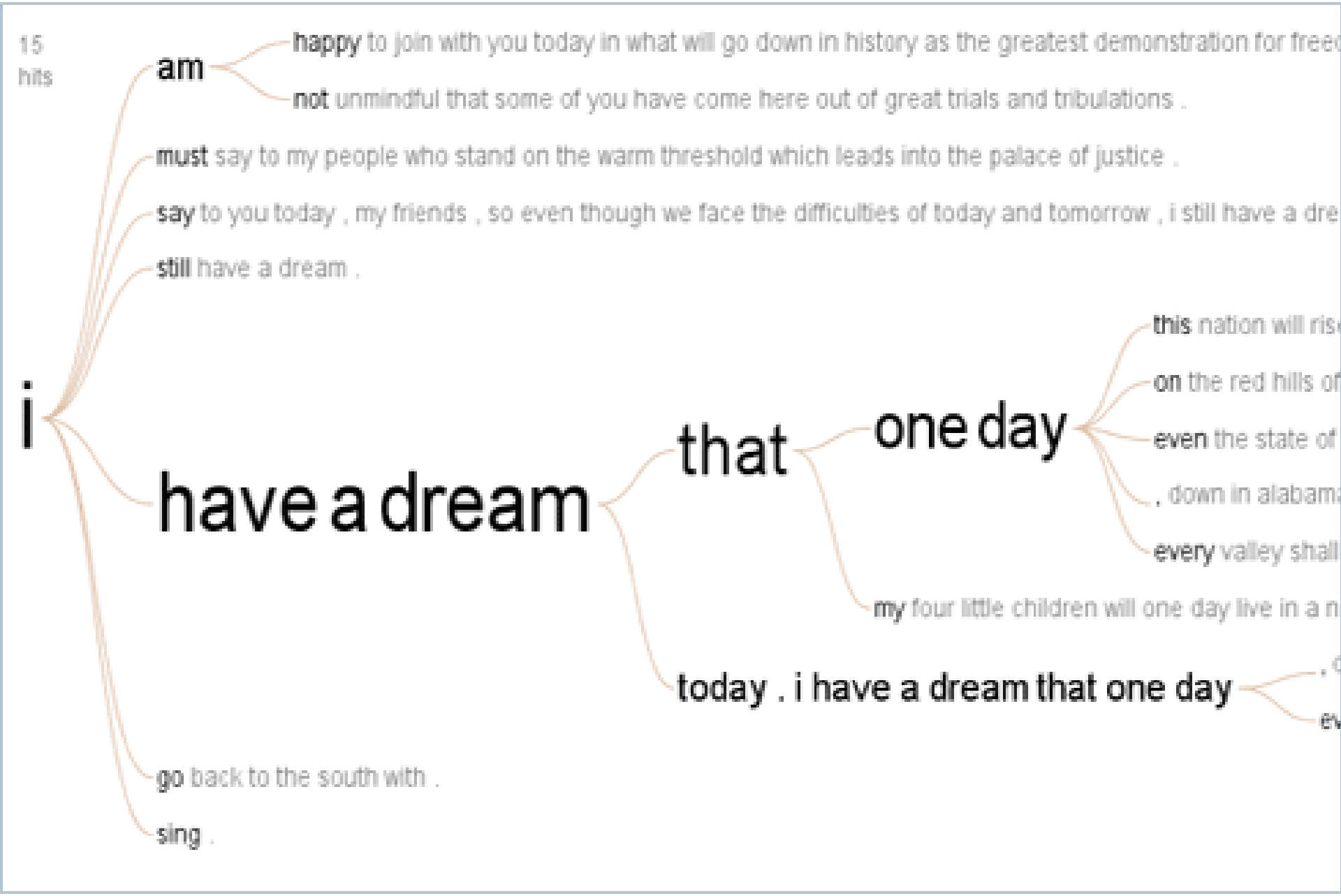


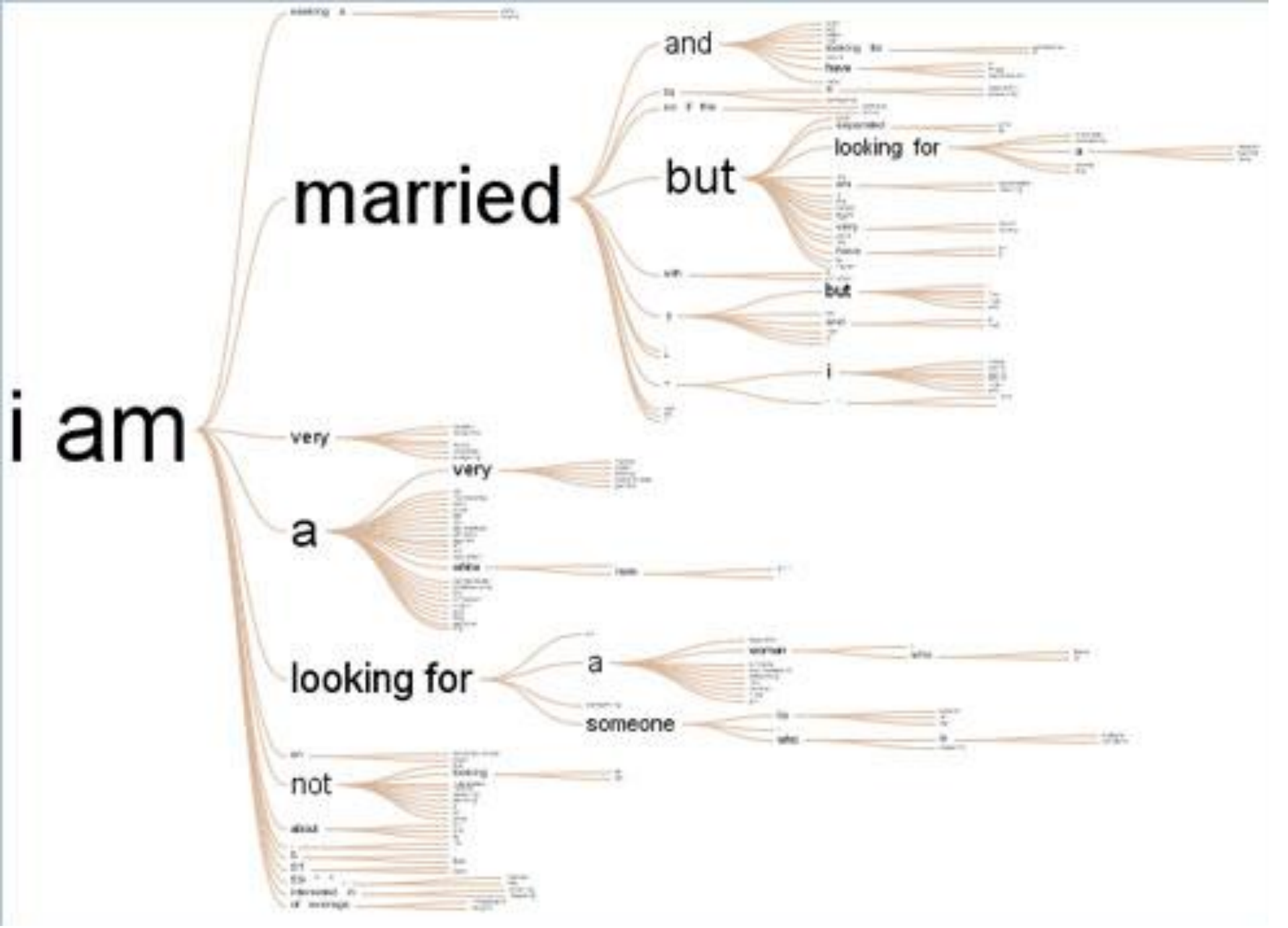
Text Mining Techniques

Text mining techniques can be understood at the processes that go into mining the text and discovering insights from it.

These text mining techniques generally employ different text mining tools and applications for their execution.

- **Information Extraction**
  - An important approach to **text mining** involves the use of natural-language **information extraction**. **Information extraction** (IE) distills structured data or knowledge from unstructured **text** by identifying references to named entities as well as stated relationships between such entities.
- **Information Retrieval**
- **Categorization**
  - **Categorization in text mining** means sorting documents into groups. Automatic document **classification** uses a combination of natural language processing (NLP) and machine learning to **categorize** customer reviews, support tickets, or any other type of **text** document based on their contents.
- **Clustering**
  - **Text clustering** is the application of **cluster analysis** to **text**-based documents.


15
hits
i
am
happy to join with you today in what will go down in history as the greatest demonstration for freed
not unmindful that some of you have come here out of great trials and tribulations .
must say to my people who stand on the warm threshold which leads into the palace of justice .
say to you today , my friends , so even though we face the difficulties of today and tomorrow , i still have a dre
still have a dream .
have a dream
that
one day
this nation will ris
on the red hills of
even the state of
, down in alabama
every valley shall
my four little children will one day live in a n
today . i have a dream that one day
go back to the south with .
sing .


i am
married
and
but
looking for
a
but
i
very
very
a
white
looking for
a
woman
white
someone
to
who
not
about