

Generating and Finding Knowledge/Information on the Web

Marian F Ursu
Department of Computing, Goldsmiths College
February 2003

Aim

- to motivate the use of meta-data from the viewpoint of the automatic processing/"reasoning" that could/should be performed by software agents (in order to reduce the amount of work that humans would do instead when looking for relevant web resources)

Overview

- human interpretations vs automatic processing/"reasoning" of web resources
 - types of automatic processing/"reasoning" tasks
- current techniques for automatic processing/"reasoning"
 - generalities
 - global search engines
 - limitations/drawbacks
- generating web resources from separate data sources
 - better structured data sources and local search engines
- meta-data

Disseminating information on the Web

- the WWW can be regarded as a medium for information dissemination
- two types of agents participate in the process of dissemination&selection:
 - human agents
 - software agents
- the ability to select relevant information increases with the degree of **refinement of the structure** of a data source
 - what is structured for humans may not be so for software agents and vice-versa

A Structured Object for Humans ...



- when displayed, a "gif" data object can have a lot (sic!) of meaning for the human observer

... but not for all software agents

- it would be tremendously difficult to develop a software that could recognise the meaning of a picture as humans do
 - apart from very restricted classes of pictures and recognition tasks, such a process is impossible nowadays

Retrieving Information on the Current Web

- term/word based search engines
- directories
- portals
 - local search engines
 - local directories

Content of a Web Page

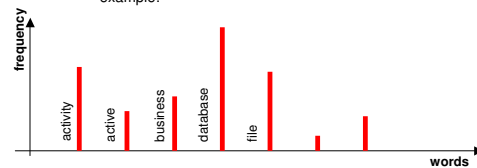
- from the point of view of a search engine
 - text objects
 - binary objects
 - pictures: gif, jpeg, tif ...
 - multimedia object (audio, movie, flash): ram, rmd, rm, avi, fla, mpeg ...
 - documents: doc, pdf, ps ...
 - recently: very limited metadata (structured text)
 - will deal with it later

Parsing Data Objects

- a search engine can only parse text
- binary objects are considered only with regards to their names

Parsing Text

- lexical profiles
 - term vector - histogram
 - usually term=word
 - not all the words of a document are considered
 - example:



- think of other possible lexical profiles

Interpreting Lexical Profiles

- rules of thumb / expert rules for assigning meaning to documents on the basis of lexical profiles
 - simple
 - compute relevance of document to a set of keywords purely on the basis of the frequency of occurrence of each keyword in the document
 - could be more sophisticated
 - match with typical profiles for different domains
 - ... imagine others ...
 - (Sousa et al, 2002)
 - may be misleading
 - the rules are not always guaranteed to work

Parsing Binary Objects

- binary objects cannot be parsed
- only their names are
 - many times these aren't "suggestive" enough
 - anyway, a file name is a **very** poor representation of its content



- possible titles: "nymphs", "water-lily", "water-lily-series-5", "nymphs-claude-monet", "ncm04660"
- what about the term-query "impressionism painting flower"?
- what about other attributes of the painting?

Global Search Engines

- Alta Vista
 - web-crawler
 - extraction of key terms (what is a key term?) + their frequency
 - for term based query, the relevant pages are ordered based on a hit factor
- Google
 - web pages are selected like in Alta Vista
 - web pages are ordered based on quotation index (how many other pages refer to the current page)
- Yahoo
 - taxonomy of terms build by humans
 - pages are classified and ordered within this taxonomy by humans

Discussion

- identify and discuss possible benefits and drawbacks of the term/word-based web search
- some drawbacks
 - translating a "highly semantic" query into a term-based query
 - irrelevant "hits"
 - "missed" resources
 - presentation
 - only list of hyperlinks
 - requires the user to browse through the listed resources
 - inappropriate ordering
 - redundancy
 - cannot integrate different resources

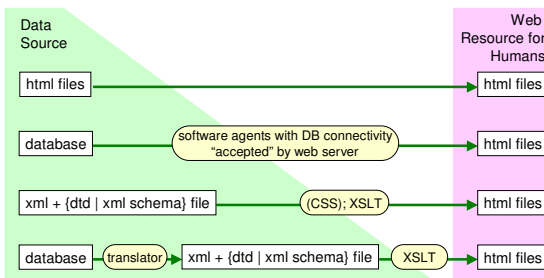
Relationship between Generation and Retrieval

- does the method of generating web resources affect the effectiveness of information retrieval?

Generating Web Resources for Humans

- web resources are traditionally intended to provide information to and be interpreted by human users only (not by software agents)
 - therefore, information/knowledge resources are/have to be translated into HTML
 - HTML is sufficiently powerful as a means of formatting/structuring audio-visual information for the human user

Generating Web Resources for Humans




Retrieving Web Resources Using Data Sources

- a database and/or an XML file may express more meaning of a data object than an HTML file
- a part of this meaning may be used by software agents (e.g. search engines)
 - it may be encoded into the agent
 - it may be expressed in a standard way, and thus possible to be understood by any software agent that adheres to the standard

Data Stored in Databases

Paintings

ID	TITLE	AUTHOR	YEAR	KEYWORDS	PICTURE
1	Nymphs	Claude Monet	1899	water lily, impressionism, lake, landscape	
2	Two Thistles	Vincent Van Gogh	1888	Arles, impressionism,
...

- the PICTURE may be sufficient for a web page
- a local search engine may use the other information stored in the database
 - more effective
 - a local search engine will encode/hard-code the structure of data
- global search engines cannot access this “extra” information
 - they have access only to the information made available in the web page

Data Stored in XML Files

- similar semantics to that of databases
- can be made directly available on the web
 - could be used by global search engines
 - a global search engine would still have to encode/hard-code the particular data structures it would work with
 - e.g., it would have to know the meaning of elements such as <title>, <author>, <year> and <keyword>
 - unfeasible, if these structures are not standard
 - alternatively, it may apply the current techniques used on text objects in HTML files to “understand” the XML elements/structures

More Meaning in Databases and XML Files

- databases and XML+{DTD|Schemas} documents express more meaning than that associated with the field/element names
 - e.g., constraints on data values
- however, this meaning is less amenable to automatic manipulations

Generating Web Resources for Software Agents

- data must be organised at least in the same manner as in databases or in XML+{DTD|Schema} documents, but its structures must be “standard”, i.e. able to be understood by any software agent
 - text objects should be structured for better automatic processing/“reasoning”
 - binary objects ought to be accompanied by meta-data to allow automatic processing/“reasoning”
 - objects should be accompanied by meta-data for better automatic processing/“reasoning”

Conclusion

- significant automatic processing/“reasoning” can be done with unstructured text and without the existence of meta-data
- however, much more can be done with structured text and meta-data descriptions/statements

References

- Fensel, Dieter, 2001. Ontologies: Silver bullet for knowledge management and electronic commerce. Springer.
- Sousa, P.A.C., 2002. A Framework for Internet Data Collection Based on Intelligent Agents – Achieved Results. KES2002 (Italy), pp. 507-511.
- Alta Vista. <http://www.altavista.com>
- Google. <http://www.google.com>
- Yahoo. <http://www.yahoo.com>
- About Search engines. <http://searchenginewatch.com>
- Deitel, H.M, Deitel, P.J, et al, 2001. XML: How to Program. Prentice-Hall.