Annals of Library and Information Studies
Vol. 61, December 2014, pp. 294-306

Use of classification in organizing and searching the web

M. P. Satija[a] and Daniel Martinez-Avila[b]

[a]UGC Emeritus Fellow, Department of Library and Information Science, Guru Nanak Dev University, Amritsar, India
[b]Graduate School of Information Science, São Paulo State University (UNESP), Marília, Brazil.

Received: 21 June 2014; Accepted: 15 November 2014

The traditional characteristics and challenges for organizing and searching information on the World Wide Web are outlined and reviewed. The classification features of two of these methods, such as Google, in the case of automated search engines, and Yahoo! Directory, in the case of subject directories are analyzed. Recent advances in the Semantic Web, particularly the growing application of ontologies and Linked Data are also reviewed. Finally, some problems and prospects related to the use of classification and indexing on the World Wide Web are discussed, emphasizing the need of rethinking the role of classification in the organization of these resources and outlining the possibilities of applying Ranganathan's facet theories of classification.

Keywords: Library Science; Information Science; Knowledge Organization; Classification; Search Engines; Semantic Web

Introduction

The Internet is one of the most wonderful, influential and popular invention of the late 20th century. Though veritably termed as the superhighway of information, it is in fact very chaotic in its information traffic. Like any other new technology its virtues are oversold[1]. It is everybody's experience on the Internet that search leads the users astray most of the times and turns it into a web of frustration. Sometimes there are jams; and other times the mad rush leads the netizens to the unintended places with numerous strangers, i.e., irrelevant hits. This is not to speak of going astray in an abandoned navigation. It is full of gold fish, but difficult to catch as our angling is primitive. The World Wide Web is unstructured where sites are constantly appearing and disappearing as it is in a constant flux. Every moment, vast amounts of uncontrolled and unorganized information is generated on it. The Web is so big that almost anything can be found on it, yet it is a challenge to find something on it. It is a matter of concern especially for the librarians whose job is to provide pinpointed and exhaustive information. It is mostly because the early designed search engines did not make use of the librarian's method to organize and search the web. There is no common classification of resources on the net, though several different methods have been followed. There can not be any comprehensive (record) catalogue of the resources available thereon, either. Current state of the Internet has been likened by Rowley and Farrow "to a library in which everyone in the community has donated a book and tossed it into the middle of the library floor"[2], and by Taylor and Joudrey "to a library where all the books have been dumped on the floor and there is no catalog"[3]. This is not a new problem though, more than ten years ago J. Rennie had aptly been quoted by Deegan saying that "at some point the Internet has to stop looking like a world's largest rummage sale. For taming this particular frontier the right people are the librarians, and not cowboys"[4]. That is to organize it using a mix of traditional and new methods of classification and indexing.

Methods of access to the World Wide Web

Before we understand how classification and indexing can be used to organize the www part of the Net, it is essential to understand the process of accessing the web. The World Wide Web, which is essentially in multimedia format, was developed at the European Particle Physics Laboratory (CERN) Switzerland in 1989. The first commercial web software was created by NEXT in 1991. It popularized the Internet so much so that web is now (though erroneously) taken synonymous with the Internet. The web is a distributed multimedia global information connecting servers on the Internet. It

merges the technique of information retrieval and hypertext which enables the users to access documents from anywhere in the world. The main characteristics of the www as described by Ellis and Vasconcelos[5] are:

- The World Wide Web organizes documents into pieces of information using the HyperText Markup Language. The HTML is a set of rules to tag and format the document.
- Each individual document or web page on the WWW has a unique address called Uniform Resource Locator (URL). Latter can be linked to other URLs or hypertext media.
- Web browser which is an interactive interface program permits to browse or navigate through the documents.
- Hypertext Transfer Protocol (HTTP) regulates communication between web browser and web server. The HTTP interprets the HTML for display and transfers the file in a natural language.

The www can be approached in two ways:

1. Direct: Ideally it is best to start with known websites. It is done easily if the URL of the site is known. Latter is its access address.
2. Indirect: Through search engines, which are web search tools.

The appearance and proliferation of search engines and indirect ways to access web pages was a natural response to the growth of information available on the World Wide Web. While knowing, memorizing and manually typing all the URLs would be a tedious and impossible task for the users, the same principles that were used for the organization of information and retrieval of documents and books in the physical environment were applied to the organization of digital information.

Search engines as indirect approaches to the WWW

A search engine is a software program designed to locate and retrieve the web documents. Started in 1994, there are a number of them now. Search engines appear, disappear and merge every year as well as some of their features and technologies do. At the date of this writing, some of the most popular and prestigious search engines include Google (https://www.google.com/), Bing (http://www.bing.-com/), Yahoo! Search (http://search.yahoo.com/), DuckDuckGo (https://duckduckgo.com/) and Baidu (http://www.baidu.com/). Of course this list might change in the forthcoming years or even months. Does anybody remember Infoseek, Altavista or MSN Search? Infoseek and Altavista became inactive (although Altavista was redirected to Yahoo! Search) and MSN Search was rebranded as Bing.

Search engines automatically create their own databases of web pages. They provide the users a search interface to enter the query expressed in natural language. They allow users to enter keywords which these match against a database. They also accept Boolean searching. Advance query formulation strategies include proximity measure, truncation, and field specific approach (name or title). The output is in the form of a list of websites with their URLs which match the query. Using embedded hyper links in the output one can reach these sites and access information provided therein. Search engines can be broadly categorized on the basis of how they search and what they search[6]:

1. Automated or statistical search engines
2. Classified directories
3. Subject specific gateways
4. Geographically restricted engines
5. Automated classified directories
6. Meta search engines

Automated or statistical search engines

Most prevalent of them are the automated search engines, also known as statistical search engines. Automated search engines are designed to locate large volumes of information as they are based on searching web pages and automatically indexing precise words from the websites. Some search engines, e.g. Excite (http://www.excite.com/), EuroFerret (http://www.euroferret.com/), however use very sophisticated computer programmes based on statistical and probability calculations and also artificial intelligence methods. These simulate human approach in identifying concepts.

Components of search engines

The components of search engines are:

The Robot: also known as spider, ant or crawler. Robots continually prowl/hunt for new documents on the Internet. Spiders wander through the web moving from website to website, reading the information on the actual site and meta tags, and following links between pages. Different search engines use different types of spiders. Some spiders are comprehensive while others are selective. Therefore, the robot of one search engine may not locate the same sites as of other one. Robots vary in methods and sophistication. Different crawlers produce totally different results. That is why a search query on different search engines produces quite different output. The robot also periodically return to the websites to check for any information that has changed, continuously updating information by eliminating dead links and adding new ones.

- The Indexer: It is a software program that automatically extracts keywords from each full text document which it adds to the index. Different search engines follow different principles. Some will index every keyword on every web page located by the spider, while others extract words from title or abstracts, and also assign relative weightings to each keyword by means of an algorithm based mostly on frequency of its occurrence. Pitfalls of such a method are obvious to librarians.

- Index: It is a store of weighted keywords which also indicates their locations. The index, essentially a large inverted file produced by a 'spider' is searchable by methods which vary according to the sophistication of the search engine concerned, but will typically include some form of Boolean query, the capacity to do a string search, or the option of limiting the scope of search to document titles. As a matter of efficiency and practicability, queries are always performed on the index, never on the whole Web 'on the fly'. The index to the database, that contains all the information extracted by the spiders at the moment of the indexing, is what users actually search when entering keywords on the search engine's interface. However, due to the lag between the indexing moment (at the last visit of the robot) and the searching moment of the user, some of the contents of the index and search results might be outdated until the robot visits the website again to update.

- The Searcher: It is a software program which compares keywords used by the net searcher with the index. Matched words are ranked in order of relevance. Searchers take users' search query and match that with the index. The result of this search is a list of sources ranked in their decreasing order of relevance to the query. Rajashekar (1999) explains[7]:

This is achieved by assigning a 'relevance score' to retrieved web pages calculated by applying a document-query similarity algorithm. These algorithms are based on statistical information retrieval techniques developed by Gerald Salton. These methods mainly depend on query-term frequencies [of occurrence] within the document across the database and with the query.

There are at least two main algorithms in a search engine: the algorithm used for searching the database index and the algorithm used for ranking the results retrieved from the database. These algorithms also vary from search engine to another as well as the results of their performance, more notoriously in the case of the ranking algorithm. While retrieving all the records that meet a specific condition in a database does not seem to be a big deal for a software, arranging for display (i.e., ranking by usefulness to the users) the thousands or perhaps millions of results that are retrieved makes the ranking algorithm one of the more valuable and distinctive aspects of the search engine. But of course the variables that affect this ranking process (and privilege some documents over others) are not fully disclosed by the search engines as they become one of their biggest commercial secrets. Some of the known aspects include search terms in the Web page, word placement and frequency (title, headings, links...) and popularity of the Web page (income links) and popularity of the Web pages linking to the Web page. However, these variables also vary from time to time to refine the performance and avoid malpractices and frauds in Search Engines Optimization (SEO). We will talk more about the ranking algorithm for the case of Google in the

next section, but, as said earlier, this ranking is not reliable.

- Metadata: It is cataloguing information regarding each source such as author, title, URL, date, keywords, etc. for resource description. Meta tags are useful to electronically access surrogates or bibliographic items lying at a certain place. But the case is different if the item itself is electronic text, which has no fixity in format or place, and that may be mirrored round the world. Its retrieval may be aided if the identification of data (metadata) is embedded in the text/document itself. Dublin Core designed by the OCLC has become a standard metadata for this purpose (Fig. 1). It also includes a place for class numbers and subject headings. This includes the use of metadata by the author of the document such as CIP data or keywords of papers in a journal. However, due to misuse and bad practices in SEO, search engines have been dismissing the possibilities of meta tags. As pointed out by Taylor and Joudrey, "by 2002, no existing search engine gave any credence to the keyword META tags found in Web documents"[3].

Subject directories

Also known as subject trees, or subject guides, these are sort of search engines designed to overcome the disadvantages of statistical search engines. They accept concept indexing and searching in contrast to word search. They use manually defined rather pre-defined metadata and employ manual classifications which are more precise in their concept based categorization and subsequent access. Editors of subject directories review sites, classify and describe them. Descriptions provide a hyper link to the web site. Yahoo! (http://dir.Yahoo!.com/), Einet Galaxy (http://www.einet.net/), Dmoz (http://www.dmoz.-org/) and INFOMINE (http://infomine.ucr.edu/) are their outstanding examples. Academic and professional directories are often created and maintained by subject experts to support the needs of researchers. Many academic subject directories contain carefully chosen and annotated lists of quality websites. INFOMINE, from the University of California, is an example of academic directory. On the other hand, directories contained on commercial portals cater to the general public and are competing for traffic. They do not restrict to sites of scholarly or research value. Yahoo! Directory would be an example of commercial directory.

Subject directories allow users to browse information by subjects such as biology, financial accountancy, digital libraries, health etc. organized hierarchically with links to different web sites. The searcher query terms are matched up with

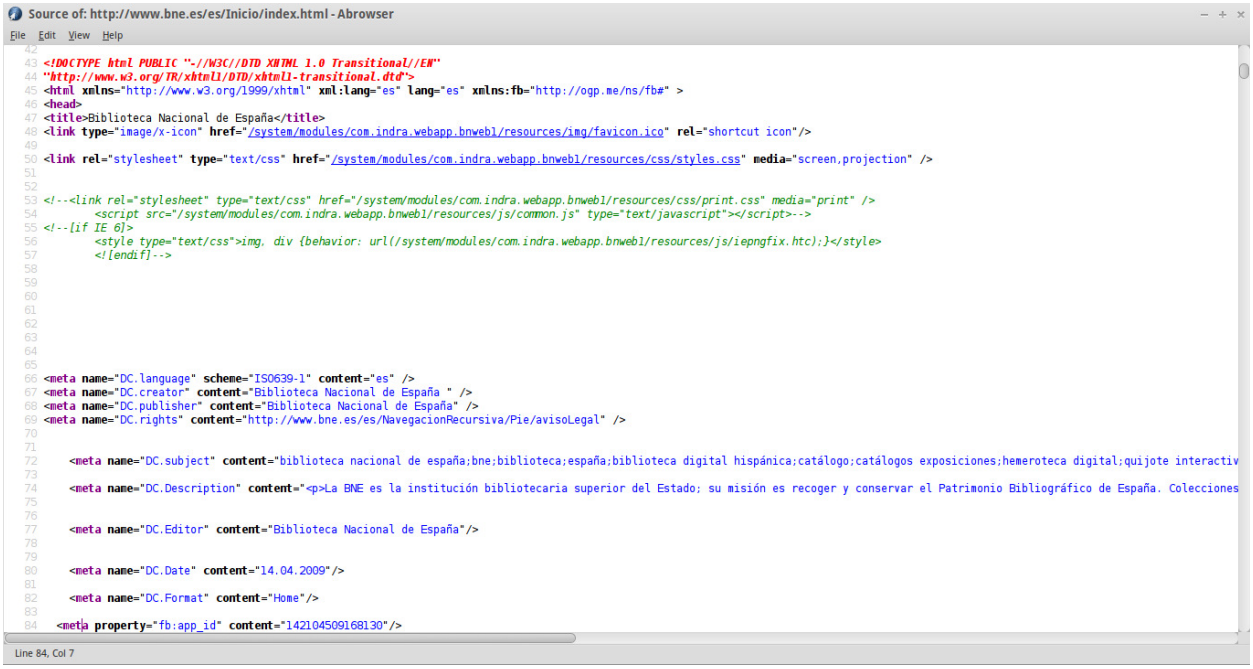


Fig. 1—Use of Dublin Core metadata in the HTML source code of a website

classification scheme and access all documents that are filed together in the index. Their advantages are obvious to the librarians. The searcher can browse the directory of categories and navigate to the relevant web sites. Such efforts are, however, primitive and do not make full use of classification as developed by traditional librarians. But this is an example of classified approach having human maintained resources. It takes time to scan and include new sites. Yahoo! the largest directory uses Robots for new sites, but employs human indexers to classify them. It is not exhaustive, but retrieves those resources which have been evaluated or value added by a professional to be worthy of inclusion. Relevance of output is high at the cost of recall. An example of subject directory using a library classification for the organization of the resources is the Subject Directory using Dewey Classification (http://education.qld.gov.au/search/-dewey.html), developed by the Department of Education, Training and Employment at the State of Queensland and, as its name indicates, using the Dewey decimal classification.

Subject specific gateways

Subject gateways are similar to subject directories but with a focus on a particular subject area. The search is narrowed to a predefined and limited subject groups. Subject gateways are usually developed by academic libraries, sometimes working in groups, or by associations/organizations interested in a subject area and dedicated to providing information on the subject. Ambiguity of terms is automatically avoided in a restricted area. Social Sciences Information Gateway SOSIG (one of the subject gateways developed within the electronic libraries programme, later Intute, currently dormant since July 2011), was an example. Other relevant examples included The Gateway to Educational Materials (also currently inactive due to the loss of funding) and DHHS Data Council: Gateway to Data and Statistics[3].

Some approaches to organize these Internet subject gateways, and other Internet resources by using library classifications have been studied in the LIS community. In 2006, Slavic gave an overview of the use of the Universal Decimal Classification (UDC) in Internet subject gateways with an English interface, such as SOSIG, OMNI or BUBL, from 1993 to 2006[8]. Kepner also studied the organization of Internet resources using the Dewey Decimal Classification (DDC). There are other scholarly articles on the use of classification schemes for the organization of Internet resources [9-11]. Finally, a very comprehensive list of known sites that used classification schemes or subject headings to organize resources was Beyond Bookmarks: Schemes for Organizing the Web (http://www.public.iastate.edu/~CYBERSTACKS/CTW.htm), compiled and maintained by Gerry McKiernan, librarian at the Iowa State University Library, although discontinued since 2001.

Geographically restricted search engines

The search area may be restricted to an area or region. Examples are Khoj, EuroFerret or UK index. Yahoo! offered this facility to restrict search to a particular country. On the other hand, while the access to some search engines might be limited in some countries (e.g., Google in China), these countries might opt for developing a country specific search engine that caters and focus on the specific local, cultural and language necessities of that country. An example of this would be the development of Baidu in China.

Automated classified directories

These search engines use a systematic classification such as DDC with automated search engines. These directories such as CyberStacks, NISS have best of both the worlds of statistical search engines and classified directories.

Meta search engines

As different search engines have different strengths, new search tools have been created. These are called multi threaded search engines which allow searchers to simultaneously search different databases using a single interface. These search engines are very fast and search through vast amount of information and pass on or broadcast the query to other search engines and provide the users with lists from other search engines. The results given in the form of URLs with hyperlinks to the respective web pages are overwhelmed with many irrelevant items. Some current examples of meta search engines are Dogpile (www.dogpile.com), GrabAll (http://www.graball.-com/), Mamma (http://mamma.com/) and Metasearch (http://metasearch.com/).

Limitations of search engines

The first limitation of search engines is that they produce huge numbers of results, many irrelevant and

yet not comprehensive. The problem is that search engines are not good at dealing with synonyms, complex concepts and meaning in context. The second limitation is that no such tool can search the entire web, not to speak of the whole Internet. The Surface Web, sometimes called the “Fixed Web” or “Public Web”, is the part of the Web that consists of fixed Web pages written in HTML or other encoded text that search engines spider and index. The information published here is not a problem for search engines. However, the Deep Web, sometimes called the "Hidden Web" or "Invisible Web", is the part of the Web that cannot be “seen”, indexed and retrieved by traditional search engines. The Deep Web includes the part of the Web that does not consist of fixed web pages, it includes the part of the Web that is served dynamically "on the fly" and more. The Deep Web is far larger than the fixed documents that many associate with the Web. Some studies have estimated that content on the deep Web may be as much as 500 times larger than the fixed Web. Generally, and although some technical limitations have been overcome such as the retrieval of PDF files, following types of documents are not retrieved, and form part of the Deep Web:

1. Information stored in databases
2. Information stored in tables (e.g., Access, Oracle, SQL Server)
3. Usually accessible only by query
4. Includes contents of library catalogs and most digital repositories
5. Multimedia files, graphical files, software, and documents in nonstandard formats
6. Web sites and intranets requiring registration or login
7. Repositories of licensed resources with restricted access
8. Dynamically created Web pages
9. Interactive tools (calculators, etc.)

Finally, some authors have also pointed out that automatic indexing and user-based retrieval systems such as Google's are not exempt from bias or subjectivity either[12,13], thus debunking the myth of search engines as a solution to the problems of bias in human indexing.

Use of classification by search engines

This paper, however limits itself to the study of the searching techniques of two very popular search engines, namely Yahoo! and Google. These are taken as examples of two diametrically opposite methods of searching. These are two major representatives of their respective class which allow searching by concepts and words respectively.

First let us see how the Yahoo! directory works. Although a very complete and interesting analysis of the Yahoo! categories, in relation to its hierarchy and navigational features, was conducted by Lee and Olson[14], for the purposes of the present study we will base our analysis on Hunter’s classification[15] From the beginning Internet search engines have recognized the fact that a sort of classification can be used for searching the World Wide Web. An illustrative extract from Yahoo!’s entry screen is given below showing the main categories from which an initial selection may be made:

Arts & Humanities:
Recreation & Sports:
Literature, Photography…
Sports, Travel, Autos, Outdoors…

Business & Economy:
Reference:
B2B, Finance, Shopping, Jobs…
Libraries, Dictionaries, Quotations

Government:
Regional:
Elections, Military, Law, Taxes…
Countries, Regions, US States…

Health:
Medicine, Diseases, Drugs, Fitness…

Earlier first level sub categories as shown above existed on the home page. But these have now been omitted there due to lack of space yielded to commercial advertisements. This home-made scheme emerged from the personal needs of its founders to keep their links organized. Clearly its categories can be mapped to any standard classification system. Each of these main categories is further subdivided into sub-categories.

Hunter[15] illustrates with a practical example: if a searcher is interested in information on ‘Compilers for

the programming language C', upon visiting the Yahoo! site a searcher would find 'Computers and the Internet' among the listed categories as shown on the home page of the website. After selecting this category, 'Programming languages' which is among the sub-categories, and selecting this will show a comprehensive list of languages: 'ABC, Active X, Ada', and so on. 'C, and C++' is one of the languages listed, and selecting this will reveal a further list that includes 'Compilers', and among the documents listed here is one entitled 'lcc: a retargetable compiler for ANSI C'. Path of this search shows that the searcher is being guided through a hierarchical structure from broader to narrower topics:

Directory > Reference > Libraries > Computers and the Internet > Programming languages > C and C++ > Compilers

Similarly a search, for, say library science faculty of the University of Mumbai, we will take up the following path:

Directory > Reference > Libraries > Library and Information Science > Education > College and University > Department and Programs > Graduate Programs > University of Mumbai > Faculty.

Glassel[16] finds this approach analogous to Ranganathan's facet analyses techniques. Other search engines provide a similar facility. It is so popular that Googling has become a verb of English language[17]. But the Google goes a different way: Google has been described as a prince of search engines. It accepts natural language query terms. Its spider called Googlebot indexes billions of pages of the web. In fact its crawler does not really roam the web – it is the other way round. Inversely, the web server returns the specified web pages which are scanned for hyperlinks. The Googlebot gives each fetched page a number for referring. Its index is an inverted file of the crawled data which lists every document that contains a given query word. Let us say, for example, we wish to search the topic Civil War. It is further assumed that the word *Civil* occurs in document numbers 5, 8, 22, 31, 56, 68 and 90, while the war occurs in 7, 8, 30, 31, 50, 68 and 92. A list of documents that contains a specified word is called a posting list. In librarians old tradition it is called uniterm indexing. When a search query is made for the topic *Civil War*, the search engine of the Google does two acts. Firstly it does post co-ordinate indexing of the posting lists to find out the documents common to both the topics of i.e. *Civil* and *War*, assuming (though not always correctly) them dealing with the topic of Civil War. In our case the common numbers from the two posting lists are 8, 31 and 68. In fact the posting lists are very huge; and librarians' traditional method both manual and mechanical may not work in such situations of huge enormity of numbers. To save time and perform the matching jobs quickly the data is stored on many computers, in fact more than five hundred. The work is divided among computers and results are consolidated. Google divides the data between many machines to find the common documents to do the job in a split second.

The second act is of relevance ranking the documents retrieved for guidance of the searcher. This ranking is of utmost importance to save the time of the searcher in sifting the relevant documents. It guides the searcher to look at or scan only the best ranked ones out of the overwhelmingly numerous links to the documents. This relevance ranking is a very ticklish issue. Two persons asking exactly the same question may consider quite different answer relevant to them[18], something that, in the case of Google, has also been criticized and named "Filter bubble"[19]. In addition, it may not be out of place to mention that in some search engines the best ranked sites are commercially sponsored. In such cases the web pages on the top have paid a certain fee to the search engine company. A huge amount of money is also involved here. Highly ranked websites get many more hits and thus the business. But this is not so with every one – only very few indulge in this unethical practice. Anyhow, ranking is always mechanical than human. The Google uses many parameters for ranking well known to librarians of information retrieval, namely:

1. Citation indexing
2. Proximity factor
3. Frequency count
4. Place of occurrence

The citation indexing is known as PageRank algorithm (also co-existing in Google with other ranking algorithms such as Google Panda, Google Penguin and Google Hummingbird). The PageRank does ranking on two aspects: Firstly, how many links there are to the retrieved web page from others

documents – that is, by how many websites it has been cited. It is citation indexing invented by Eugene Garfield. The other is the quality of the website citing that document. The quality being subjective is assessed manually. The links from quality websites are given a higher rank than the less reputed ones. Even two links from a reputed websites such as Time.com are given higher ranking than the say five links from a less reputed one. The other consideration is the proximity factor. A document containing words *Civil* and *War* next to one another is considered more relevant and thus given higher ranking than a document containing these words occurring in isolation or separated by many other words. Third factor in ranking is the frequency count. If the words *Civil* and *War* occur several times in a document that is considered more likely about the topic as compared to the document where it occurs say once or twice. But the problems and pitfalls of this blind frequency count are well known, especially in free-text searching. A concept may be expressed by many names – even a concept may be dealt without naming it! It is apart from ironical expressions where the words mean just the opposite. Finally there is the place of occurrence. Apart from the number of occurrences, the place from where a word comes is quite important. If the query word occurs in the title, subject heading or as feature or section heading, then it is justifiably given higher rank than other documents containing the same word at other places in the body of the text.

The overall ranking is a combination of all these factors. The documents with the relatively higher score are considered as best matches or most relevant to the query. Further, to help ascertain relevance of the ranked output the list also shows a clipping of the pages showing the use of the word and its place of occurrence. The output shows the ranked URLs with hyperlinks and page clippings in a split second. It may be warned that being mechanical this ranking is not always reliable, even if not commercially sponsored.

Word searching has its own inherent problems. Google search engine is hamstrung by poor scholarly search methods. As mentioned before, one of its problems is that they are not able to recognize the conceptually equal words or synonyms or foreign languages. "They do not allow you to recognise related sources whose term you cannot think of before hand"[20]. Further, relevance rank is not the same as conceptual categorisation – latter cannot be done by a machine algorithm. This categorization is crucial to scholarship.

Among Google's search facilities is a search by category. It is claimed that it provides a convenient way to refine the search on a particular topic. Searching within a category of interest allows to quickly narrow down to only the most relevant information. There are two ways of doing this:

1. One is to work down through the hierarchies in a similar manner to the described previously. For instance for information on "potted plants", one might choose 'Home' from the top display of categories, then 'Gardens' from the display of sub-categories, then 'Plants' from the categories listed under 'Garden'.

   Home → Garden → Plants → Potted

2. The other way is to search first for a particular topic and then narrow the search by selecting particular category of interest. It is just like operating a relative index of the Dewey Decimal Classification. For example, if a search is conducted for 'Venus', the result will be a wide range of sites covering such as Venus as a planet, Venus as a Goddess, Venus temple, the Venus Dating Agency, Venus Internet Ltd., Radio Venus, and so on. If hierarchy of categories given against particular sites is examined, for example:

   Science > Astronomy > Planets

   Or

   Literature > English > Plays > Elizabethan > Shakespeare > Work

It is comparatively easy to select the category into which specific requirement falls. Clicking 'Elizabethan literature' from the last hierarchy above, for example, will lead to those sites concerned with Shakespeare. It is reverse of the disciplinary approach and similar to the one as used by J D Brown (1862 – 1914) in his Subject Classification (1906).

The semantic web, the growing application of ontologies and linked data

In the recent years, some voices have also pointed out that the solution to these problems might aim to the semantic web, the growing application of

ontologies and linked data. According to Tim Berners-Lee, inventor of the WWW, the Semantic Web "is a web of data, in some ways like a global database"[21]. In 2001, he also extended this definition stating that "the Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation"[22]. The Semantic Web is a more highly structured version of the Web intended to allow intelligent robots to merge information from diverse sources. It is a collection of information in a structured, machine readable format. It assumes that all knowledge can be written as a relationship between two or more items. The main characteristics of the semantic web presented by Berners-Lee are:

- Expressing meaning, that is bringing structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users without discriminating among types of information, language, cultures, etc.

- Knowledge representation, for the semantic web to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning. To achieve this, two main technologies were proposed for the development of the semantic web: the eXtensible Markup Language (XML), to allow people to create their own tags and add an arbitrary structure to the documents, and the Resource Description Framework (RDF), to express meaning and encode it in sets of triples (subject, predicate and object), that can be written using XML tags.

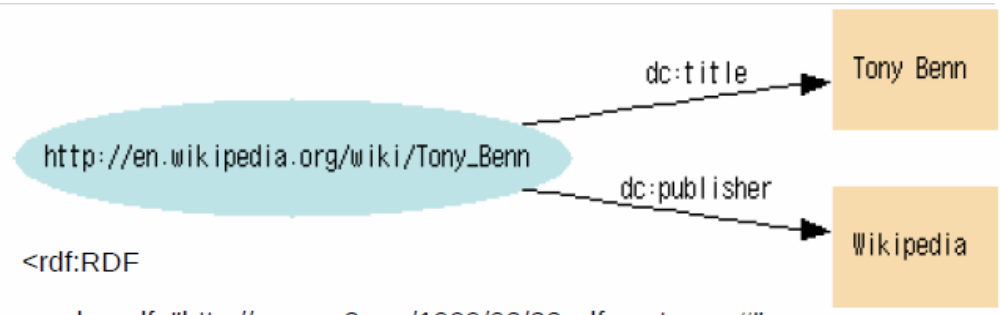
- Ontologies, that are documents or files that formally define the relations among terms, typically having a taxonomy and a set of inference rules. Here, ontologies can be used to deal with problems related to terminology and ambiguity.

- Agents, that are programs that collect Web content from different sources, process the information and exchange the results with other programs. To increase their effectiveness, machine-readable Web content and automated services should include semantics.

- Evolution of knowledge, and here it is said that in naming every concept simply by Uniform Resource Identifiers (URIs), anyone can express new concepts that they invent with minimal effort and link these concepts into a universal Web.

RDF: Resource Description Framework

RDF was developed by the World Wide Web Consortium (W3C) in 1999. RDF is a language for representing information about resources in the WWW. It is particularly intended for representing metadata about Web resources, such as the title, author, and modification date. RDF can also be used to represent information about things that can be identified on the Web, even when they cannot be directly retrieved on the Web. RDF is intended for situations in which this information needs to be processed by applications, rather than being only displayed to people. RDF provides a common framework for expressing this information so it can be exchanged between applications without loss of meaning. In order for this model to be useful, it must be expressed concretely, that means that RDF may be encoded in XML or in some other markup language. RDF-structured metadata enables the exchange and re-use of metadata in ways that are semantically unambiguous. RDF model is based on the idea of making structured information statements in the form of subject-predicate-object expressions (RDF triples). The subject of an RDF triple represents the resource. The predicate represents traits, characteristics or aspects of the resource and expresses a relationship between the subject and the object. Figure 2 is an example of knowledge representation using RDF and XML, taken from Wikimedia Commons. In this example, it is stated in a machine understandable language that the resource to be described (the subject of the statement) is http://en.wikipedia.org/wiki/-Toni_Benn. The title, according to the definition of title in Dublin Core 1.1 (a predicate), is *Tony Benn* (an object). The publisher (another predicate) is *Wikipedia* (another object). This information can be exported and shared in a form that others can understand using the XML format. In the XML file, it is stated, also in a machine understandable language, that the file will use RDF and the Dublin Core metadata schema to specify the meaning of the tags. The exact definition and meaning of these schemas are specified in the referred XML namespaces denoted by xmlns. The advantage of this




```
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description rdf:about="http://en.wikipedia.org/wiki/Tony_Benn">
   <dc:title>Tony Benn</dc:title>
   <dc:publisher>Wikipedia</dc:publisher>
  </rdf:Description>
</rdf:RDF>
```

Source: http://commons.wikimedia.org/wiki/File:Rdf-graph-example-TonyBenn.png

Fig. 2—Example of RDF and XML

representation is that machines are able to process this information and treat it in a way that enhances its use and retrieval automatically.

Ontologies

Ontologies are hierarchical relationships developed in computer science to attempt to formalize abstract concepts and relationships between them for artificial intelligence research. They are models of the entities that exist in a certain domain and the relationships among them. While the ability to create or discern patterns and connections between items is a basic property of the brain, ontologies try to provide patterns so a machine can make new connections between concepts. Ontologies have the following parts:

- Individuals: instances or objects (the basic building blocks of an ontology).
- Attributes: properties, features, characteristics or parameters that objects can have.
- Classes: sets of objects that are related in some way.
- Relationships are ways in which classes and objects are related to one another

The most typical kind of ontology for the Web has a taxonomy and a set of inference rules. The taxonomy defines classes of objects and relations among them and the inference rules allow computers to make artificially-intelligent connections among concepts. The meaning of terms or XML codes used on a Web page can be defined by pointers from the page to an ontology expressed in an ontology language such as OWL (Web Ontology Language).

URIs: Uniform Resource Identifiers

A key element of the semantic web is to identify things and relationships in a way that can be understood by machines. Every ‘thing’ has to have an identifier that distinguishes it from any other thing. The URI (Uniform Resource Identifier) is the identifier of those things that is used in the statements of the semantic web. The common URL (Uniform Resource Locator) (e.g., http://en.wikipedia.org/wiki/-Toni_Benn) is in URI format and is the preferred identifier to use in the semantic web. The URI is what

allows a resource to be a thing on the Web and be actionable on the Web. If somebody wants to link to something it has to have a URI. If someone wants to locate it, it has to have a URL (which is also a URI). URLs therefore can be used as identifiers as well as locators.

In addition to URIs used as resources (subjects) in RDF triples (e.g., http://en.wikipedia.org/wiki/-Toni_Benn), URIs can also be used as properties (predicates) (e.g., http://purl.org/dc/elements/1.1/-subject -the Dublin Core specification for subject) and values (objects) (e.g., http://id.loc.gov/authorities/-classification/Z7234.C3.html). There are several subject URIs published by authorized organizations that can be used to unambiguously state subject headings and classification numbers. Examples of these subject URIs are the Linked Data Service that provides access to commonly found standards and vocabularies promulgated by the Library of Congress (http://id.loc.gov/) and the Dewey Decimal Classification linkable data (http://dewey.info/).

Linked (Open) Data

The semantic web uses linked data to connect information that was not previously connected. It allows querying data as in a database instead of current search engine's text-string matching and relevance ranking algorithms. As Berners-Lee put it: "The semantic web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data"[23]. For this, Berners-Lee provided four rules:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names, and here Berners-Lee emphasizes that HTTP URIs are names (not addresses).
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL), that is to make it available and in a way that can be understood and processed by machines and others
4. Include links to other URIs, so that they can discover more things

In 2010, Tim Berners-Lee added the need of using legal mechanisms such as the GPL-derived licenses (such as Creative Commons) to guarantee the free use of data, stating that "Linked Open Data (LOD) is linked data which is released under an open license, which does not impede its reuse for free."

The definitions of linked data and linked open data have brought fresh air to the semantic web and its technologies such as RDF, OWL and SKOS. Linked open data has also inspired some new and promising projects such as Linked Open Vocabularies (LOV), aimed to promote and provide access to vocabularies in the "cloud"; schema.org, a project developed by Google, Microsoft and Yahoo to provide Web publishers with a universal vocabulary to describe their pages using linked data; and HIVE (Helping Interdisciplinary Vocabulary Engineering), that is both a model and a system that supports automatic metadata generation by drawing descriptors from multiple Simple Knowledge Organization System (SKOS)-encoded controlled vocabularies[24].

Problems and Prospects

However, many of these projects and technologies remain underused and perverted by the commercial interests that are sometimes leading the organization of resources on the WWW (as in opposition to the professional organization and classification by librarians). This view, written by Robert Cairo, was quoted at the Third International Conference on Intelligent Systems Modelling and Simulation in 2012, and it says that: "The Semantic Web will never work because it depends on businesses working together, on them cooperating. We are talking about the most conservative bunch of people in the world, people who believe in greed and cut-throat business ethics, people who would steal one another's property if it weren't nailed down. The people who designed the semantic web never read their epistemology (the part of philosophy that is about the study of how we know things) texts. But the big problem is they believed everyone would work together: - would agree on web standards - would adopt a common vocabulary - would reliably expose their APIs so anyone could use them"[25].

On the other hand, Dodd[26] was quite right to describe some of these attempts at categorization as 'semi-professional'. Although the hierarchical structures do support subject browsing, the nature of the 'classification' in these search engines does not appear to be as systematic as can be found in the more

traditional library schemes; frequently cross-classification is apparent, and a further disadvantage is that they do not have a notation. Though a notation is not absolutely necessary for a system of classification to work, it can provide added value and save us from the problems of uncontrolled vocabulary.

Organization of electronic and internet resources is needed without doubt. A resource is not a resource if it is not organized and controlled. All the books in the world won't help you if they are just piled up in a heap, aptly says Eddings[27]. Question is how far these traditional systems and skills are effectively transferable in the new information environment. So far use of classification in databases and Internet sites is shallow and too simple to make way for useful browsing. These may be seen as, to quote St. Paul, "having form of classification but denying its power"[28]. The need is as Newton puts, to develop traditional classifications and thesauri "to provide a syndetic structure which can be used as a basis for systematizing hypertext linkages between electronic documents". We need to turn classification into knowledge organization tool for efficient retrieval of material for virtual library[29].

The hypertext documents cannot adequately be classified by a traditional classification with static hierarchies. Internet flexibility or manipulation of documents should allow for creation of more linkages between subject hierarchies. Even cross classification, a defect in traditional systems can be a virtue here. Classification theory can provide many new avenues. But one agrees with Newton that "Ranganathan's distinctive and radical thinking in his numerous books and papers and the significance of the paradigm shift in classification theory which he inaugurated has not yet been fully mined"[29]. Ranganthan's facet analysis is immensely helpful in query formulation for better precision of output. The WWW a huge, turbulent and complex source of information will definitely benefit from traditional skills and tools of libraries. Though the library professionals have been quite successful in applying library classification to networked resources, yet much remains to be done to make them more amenable to the new environment.

Conclusion

With the proliferation of information on the World Wide Web there was also a new necessity of organization. Several approaches have been taken to organize that information and obviously some of them came from librarians and information scientists. However, as in the case of brick and mortar libraries, not all systems and approaches have been proved to be equally adequate or successful. As in the case of library classifications, we think that the application of Ranganthan's facet analysis to the classification of information on the World Wide Web would be a great contribution with the potential to overcome some of its historical problems.

References

1. Stockwell F A, *History of information storage and retrieval*, (McFarland & Co.; Jefferson), 2001.
2. Rowley J E and Farrow J, Organizing knowledge: An introduction to managing access to information, 3rd ed; 2000 (Aldershot:Gower)
3. Taylor A G and Joudrey D N, *The Organization of information*, 3rd edn, (Libraries Unlimited; Westporet), 2009.
4. Deegan M, The spectrum of digital objects in the library and beyond, In GE Gorman, ed., *Digital factors in library and information services, (Facet; London), 2002, p. 21.*
5. Ellis D and Vasconcelos A, Ranganathan and the Net: Using facet analysis to search and organize the World Wide Web, *Aslib Proceedings*, 51 (1) (1999), 3-11.
6. Gash S, *Effective literature searching for research*, 2nd edn, (Gower; Aldershot), 2000.
7. Rajashekar T B, Internet and Web search engines, *Library Herald*, 37 (1) (1999), 60-74.
8. Slavic A, UDC in subject gateways: experiment or opportunity? , *Knowledge Organization*, 33 (2) (2006), 67-85.
9. Vizine-Goetz D, Classification schemes for Internet resources revisited, *Journal of Internet Cataloging*, 5 (4) (2002), 5-18.
10. Vizine–Goetz D, Using library classification for Internet resources. In OCLC Internet Cataloging Project Colloquium, 1999. Available at: http://staff.oclc.org/~vizine/Intercat/-vizine-goetz.htm (Accessed on 10 Jun 2014).
11. Zins C, Models for classifying internet resources, *Knowledge Organization*, 29 (1) (2002), 20-28.
12. Segev E, *Google and the digital divide: the biases of online knowledge*, (Chandos; Cambridge), 2009.
13. Hjørland B, User-based and cognitive approaches to knowledge organization, *Knowledge Organization,* 40 (1) (2013), 11-27.
14. Lee, H-L and Olson H A, Hierarchical navigation: an exploration of Yahoo!! Directories, *Knowledge Organization*, 32 (1) (2005), 10-24.
15. Hunter E J, *Classification made simple*, 3rd edn, (Ashgate; Burlington), 2009.
16. Glassel A , *Was Ranganathan a Yahoo!!?* (1998). Available at: https://scout.wisc.edu/scout/toolkit/enduser/archive/-1998/euc-9803 (Accessed on 10 Jun 2014).

17. Harvey R and Hider P, *Organising knowledge in a global society*, (Charles Sturt University; Wagga), 2004.

18. Notess G R, The never-ending quest: Search engine relevance, *Online* 24 (3) (2000), 35-40.

19. Pariser E, *Beware online "filter bubbles"* (2011). Available at: http://www.ted.com/talks/lang/en/eli_pariser_beware_-online_filter_bubbles.html (Accessed on 10 Jun 2014).

20. Mann T, *Will Google's keyword searching eliminate the need for LC cataloging and classification* (2005). Available at: www.guild2910.org/searching.htm (Accessed on 10 Jun 2014).

21. Berners-Lee T, *Semantic Web road map* (1998). Available at: http://www.w3.org/DesignIssues/Semantic.html (Accessed on 08 Jun 2014).

22. Berners-Lee T, Hendler J and Lassila O, The Semantic Web, *Scientific American*, May (2001), 1-4.

23. Berners-Lee T, *Linked Data* (2006). Available at: http://www.w3.org/DesignIssues/LinkedData.html Accessed on 08 Jun 2014).

24. Greenberg J, Losee R, Pérez Agüera J R, Scherle R, White H and Willis C 2011, HIVE: Helping Interdisciplinary Vocabulary Engineering, Bulletin of the American Society for Information Science and Technology, 37 (4) (2011). Available at: http://www.asis.org/Bulletin/Apr-11/Apr-May11_Greenberg_etAl.html (Accessed on 08 June 2014).

25. Agarwal P R, Semantic Web in comparison to Web 2.0, in *Third International Conference on Intelligent Systems, Modelling and Simulation (ISMS), 2012, (Jyoti Microsyst. Pvt. Ltd.; Ahmadabad), p. 558-563.*

26. Dodd D G, Grass root cataloging and classification: Food for thought from World Wide Web subject- oriented hierarchical lists, Library Resources and Technical Services, 40 (3) (1996), 275-286.

27. Eddings D, *King of the Murgos* (Ballantine Books; New York), 1987.

28. MacLennan A, Classification and the Internet, In *The future of classification*, R Marcella and A Maltby, eds., (Gower; Aldershot), 2000, p. 59-67.

29. Newton R, Information technology and new directions, In *The future of classification*, R Marcella and A Maltby, eds., (Gower; Aldershot), 2000.