Developing Semantic Applications for INDEST e-resources

M. Ratnasamy
Central Library
I.I.T., Kharagpur

- Educational Institutions:
- Higher Education System of a country plays a significant role in the creation of skilled human resources.
- According to the Swami Vivekananda, we want that education by which “Character is formed, strength of mind is increased, intellect is expanded and by which one can stand on one’s own feet”.
- In India the total students enrollment was 228 millions in 12,21,198 all types of educational institutions. Out of these, 12,00,000 Schools, 20,782 Colleges, 416 Universities, 62,10,000 teachers (6.21 million).
-
- [UGC News Letter, July 2009]

- INDEST-AICTE Consortium:
- About 918 Institutions have enrolled as Members in INDEST-AICTE Consortium.
- There are about 24 major e-resources such as ACM, Science Direct, IEEE, Springer, ABI/Inform, SciFinder Scholar, Web of Science etc., being accessed.
- About 12676 full-text journals are being made available for access through the Consortium depending upon on the type/level of the Institution. This also includes about 2376 Open Access full-text journals.

- [INDEST Web Site accessed on 14-01-2010]

- World Wide Web:
- The Sheer Size of the web has led to a situation where simple statistics about it are unknown, for example, its size or the percentage of pages in a certain language.
- We can use, for example, random URLs to estimate the distribution of the length of web pages, the fraction of documents in various Internet Domains and on the fraction of documents written in various languages.
- We can also determine the fraction of web pages indexed by various search engines by testing the engines for the presence of pages chosen uniformly at random.
-
- [M.Henzinger and S.Lawrence, 2004]

- IP Address Sampling:
- There are ~4.3 billion IP addresses.
- Of the 4.3 billion possible IP addresses, some are unavailable and some are known to be unassigned.
- The distribution of server types found from sampling 3.6 million IP Addresses, about ~83% of servers were commercial.
- Where as, about ~6% of web servers were found to have scientific/educational content (defined here as university, college and research laboratory servers).

- [M.Henzinger and S.Lawrence, 2004]

- Metadata:
- The following was observed while analyzing the metadata usage on the homepages of each server:
- Only 34.2% of servers contained the common “keywords” or “description” metadata tags on their home page.
- Low usage of the simple HTML metadata standard suggests that acceptance and widespread use of more complex standards, such as XML or Dublin Core.
- About 0.3% of sites contained metadata using the Dublin Core Standard.
- High diversity was also noted in the HTML metadata tags found, with 123 distinct tags, suggesting a lack of standardization in usage.

- [M.Henzinger and S.Lawrence, 2004]

Thesaurus, taxonomy and ontology

- Looking at the applications of thesauri, taxonomies and ontologies it is easy to see a progression of ideas that has resulted in some overlapping of details.
- The thesaurus has been the domain of information scientists;
- Taxonomies appear to have been generated by a combination of information technologists and system developers in corporate business;
- The ontologies have been adapted from the work of philosophers by working in Artificial Intelligence.
- The fundamental components of ontologies may include concepts in the form of nouns, adjectives, verbs, phrases, morphological variations, notes on concept relationships and usage.
- Berners-Lee defines “Ontologies as “ a document or file that formally defines the relations among terms. The ….ontology for the web has a taxonomy and a set of inference rules.”
- [Alan Gilchrist, 2003]

- Users unanticipated conceptual framework – Facet Analysis
- In order to accommodate the unanticipated conceptual framework within which patrons might encapsulate their research requirements, Ranganathan developed, what he called, the ‘faceted’ approach to knowledge classification.
- For using faceted analysis, it is no necessary to have a completely comprehensive understanding of the entities housed in a digital repository or the relationships between the ontological terms used to describe the entities.
- Facet Analysis helps to recognize the Basic Subject and each of the Isolates of a subject and arranging them in a preferred sequence in accordance with prescribed rules.
- [Robert Fox, 2005]

- Facet Analysis is work done in the Idea Plane and in the Verbal Plane (basic subject, terms, ..).
- It does not involve any work in the Notational Plane. In fact it does not require any notational system [DC, UDC, URI].
- It does not depend on any scheme of Classification.
- The standard procedure of the Facet Analysis helps to determine the kernel terms needed in choosing and specifying the name of the subject of a given document and the sequence of rendering its name in the heading.
- The standard procedure has four steps yielding successively the full Title, the kernel title, the analyzed title and the transformed title respectively of a document. [S.R.Ranganathan, 1964]

- e-Science;
- Real Scientific Progress depends on collaboration and making connections between ideas, people and data.
- No one scientific laboratory has the resources or tools, the raw data or derived understanding or the expertise to harness the knowledge available to a scientific community.
- e-Science is science performed through distributed global collaborations between scientists and their resources enabled by electronic means in order to solve scientific problems.
- U.K. has earmarked 240 million pounds to develop a sustainable and effective e-Science e-infrastructure.
- The web-based distributed information infrastructure facilitates the scientists to:
- search the web for content; interpret and process the content of web pages; infer the cross-links between information; integrate the content from multiple resources. [Carole Goble, 2005]

- **‘Adam’ Robot Scientist**
- Scientific knowledge is best expressed in formal logical languages.
- Only formal languages provide sufficient Semantic Clarity to ensure reproducibility and free exchange of scientific knowledge.
- Despite the advantages of logic, most scientific knowledge is expressed only in natural languages.
- This is now changing through developments such as Semantic Web and Ontologies.
- ‘Adam’, the Robot Scientist has autonomously generated functional genomics hypotheses about the yeast “Saccharomyces Cerevisiae” and experimentally tested these hypotheses by using laboratory automation.
- This has been possible due to the formalization of scientific investigation in logic.
- **For the formalization they have used the ontology of scientific experiments viz EXPO/KABORS**
- **[R D King etal, 2009]**

- Current web technology comes up short when it comes to supporting the needs of the collaborative and interdisciplinary “e-Science”.
- New Web technologies are emerging with the potential of revolutionize the ability of scientists to do collaborative work;
- Semantic Web [Berners-Lee, 1999]] is designed :
- to improve communications between people using differing terminologies;
- to extend the interoperability of databases;
- to provide tools for interacting with multimedia collections; and
- to facilitate “agent-based” computing in which people and machines work more interactively.
- Semantic Web uses web languages RDF, RDFS, OWL etc which go beyond the presentation capabilities of HTML and the document-tagging capabilities of the XML. [James Hendler, 2003]

- Metcalfe’s Law and IP Address
- James Hendler and Jennifer Golbeck have outlined the impact of Network Effect on the performance of Web 2.0 and of Semantic Web.
- Bob Metcalfe in 1980s hypothesized that while the cost of the network grew linearly with the number of connections, the **value** was proportional to the square of the number of users.
- For example, given, n users of ethernet cards, the number of possible connections that can be made is
- n(n-1) = O (n*n)
- The tremendous growth of sites such as MySpace, Facebook and YouTube indicates that the Social Networking construct is critical to the success of Web 2.0 applications.
- The fact that sharing of content can be enhanced by personal connections, rather than, via search or other techniques has emerged as a major and perhaps defining aspect of successful Web 2.0 applications
- [James Hendler and Jennifer Godbeck, 2008]

- Semantic web or Web 3.0
- Semantic Web attempts to mine meaning from the words/terms in the web content. It uses web languages to express the relationships between terms where these terms are assigned specific URIs.
- Some of the most used Semantic Web vocabularies like the Friend of a Friend [FOAF] ontology get their primary value not from the terms they express, but from the many instances linked to each other through the common (and unambiguous) vocabulary.
- While inferencing is an important aspect of web, **the ability for terms to be linked is a critical difference** between **RDF** based languages and earlier **Knowledge Representation** languages.
- Couples with languages such as SPARQL and RDFa which provide a technology base for making Semantic web applications interoperate more smoothly with traditional web applications.
- [James Hendler and Jennifer Godbeck, 2008]

- Facet analysis – Semantic Web – IR Techniques – Indexing
- The result of Facet Analysis is a controlled vocabulary of simple concept terms organized into facets for pre-coordinate or post-coordinate operation. This provides flexibility for both the indexer and the searcher.
- The concept terms should be derived from and reflect the actual terms used in the field rather than being used on some priori conceptualization of the subject (620.1).
- The order of terms within facets and sub-facets should be decided in terms of usefulness or naturalness to the user.
- Indexing - Indexing is one of the most widely used formal devices for organizing knowledge for information retrieval.
- There are two main types of methods for indexing or searching for documents on the Web:
- a) Word based indexing ; and b) Concept based indexing.
- [David Ellis and Ana Vasconcelos, 1999]

- Word based indexing: Words in the source document are extracted regardless of their meaning and these words are used as indexed. Some systems include statistical devices to assign weights to each term and therefore determine which are the most important terms in the document by frequency of occurrence. It is then possible to select for representation only the most frequently occurring terms.
- Search Engines on the WWW work through automatic word indexing of Websites.
- Concept based indexing: Concepts are ideas of things, to be distinguished from names of things. “The creation of Concepts in subject indexing is based on the association of representations of objects, ideas, processes and other entries with related data which pre-exists in the mind”[Fosket21].
- So Concept based indexing is contrary to word based indexing. Concept based indexing require the identification of the concepts which are represented by the terms used in the documents, rather than just extracting terms [words] used in the document. [David Ellis and Ana Vasconcelos, 1999]

- It is known that there is total lack of control over the creation of www materials. Also there is general lack of proximity between searcher and indexer.
- Web Semantics: Semantic Web is more a social rather than a technological problem. We need to help the process of adding semantics by:
- i) by extracting appropriate concepts from the web content;
- ii) by extracting relevant terms from the link structure; and
- iii) by identifying relevant and related concepts from the users' queries
- [Baeza-Yates, 2008]

- Proposal
- Let us identify the specific unique mandate of the members of INDEST-AICTE Consortium.
- Let us conceptually extract the specific subject areas covered by the 12500 full-text journals.
- Let us group these journals by conceptual coverage and link these journals so as to effectively promote the mandate of the consortium members by using Semantic Web languages and Technology.
- Facet Analysis approach may effectively facilitate to achieve the above objectives.
- Let us revisit Ranganathan’s ideas for the greater benefit of library users and mankind.

- References:
- 1. M. Henzinger and S. Lawrence. Extracting Knowledge from the World Wide Web. Proc. National Academy of Sciences. Vol. 101(2004), 6th April 2004, pp 5186-5191.
- 2. Alan Gilchrist. Thesauri, taxonomies and ontologies – an etymological note. Journal of Documentation, V.59(2003), No.1, pp 7-18.
- 3. Robert Fox. Cataloguing our information architecture. OCLC Systems and Services: International Digital Library Perspectives. Vol.21(2005), No.1, pp23-29.
- 4.S.R.Ranganathan. Subject Heading and Facet Analysis. Journal of Documentation, Vol.20(1964), N.3, pp109-119.
- 5. Carole Goble. Using the Semantic Web for e-Science: Inspiration, Incubation, Irritation.

  In: Y.Gil et al(Eds): ISWC 2005, LNCS 3729, pp1-3, Springer Verlag, 2005.

6. R.D. King et al: The Automation of Science. Science, Vol.324, 3rd April 2009, pp 85-89.

7. James Hendler. Science and the Semantic Web. Science, vol.299, 24 Jan. 2003, pp520-521.

8. James Hendler and Jennifer Golbeck. Metcalfe's Law, Web 2.0 and the Semantic Web. Web Semantic: Science, Service and Agents on the World wide Web, Vol.6(2008),pp14-20.

9. David Ellis and Ana Vasconcelos. Ranganathan and the Net: using facet analysis to search and organize the world wide web. Aslib Proceedings, Vol.51(1999), pp. 3-10.

10. R. Baeza-Yates. From capturing semantics to semantic search: a virtuous cycle. In: S.Bechhofer et al (Eds) : ESWC 2008, LNCS 5021,pp1-2, Springer-Verlag, 2008.

- Thank You