Information retrieval meets semantic web

Juan Bernabé Moreno
University of Granada, Department of Computer Science and Artificial Intelligence

Abstract— **Google is the most prominent proof that the first usage of the WWW user is the search for information. But the Information Retrieval was born much before the WWW and with very successful results. That means only that the fundamentals should be better explained and better understood, which will allow us to present the idea of semantic web and to better highlight the advantages it can bring to the Content Management Systems. That's exactly what this work takes up.**

***Index Terms*— Information retrieval, semantic web, CMS**

I. INTRODUCTION

When we talk about content management system we refer to systems that take care of the data storing and structuring of web documents and their presentation over the internet. Condition sine qua non for the significance of the content management systems is the existence of the information consumers, and it can make us think about the way the information is retrieved and if there's room for improvement.

This work takes up the interfaces between both disciplines, the area where the information consumers are faced on their way to the Content Management Systems functionalities and contents. We will explain how the upcoming CMSs contain features that are already since several decades under research of Information Retrieval (see Figure 1)

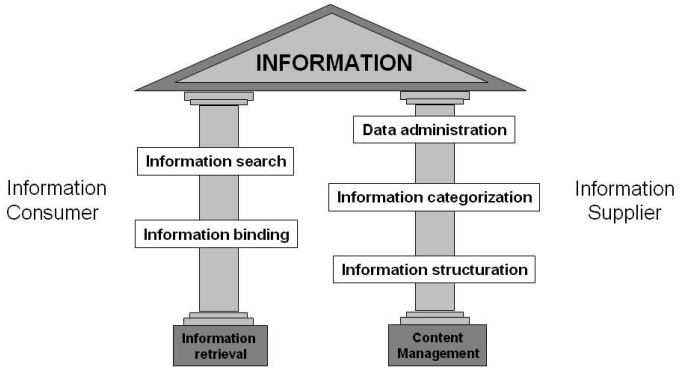

Figure 1

II. INFORMATION RETRIEVAL

Already in 1974, Information retrieval was defined by ISO-Standard ISO 2382/1 as the "Actions, methods and procedures for recovering stored data to provide information on a given subject" [1] This definition clearly points out that IR is not a mere data querying but refers to how the information is related to a particular topic (and exactly that is the great challenge)

For example, the finding of key-word in a text (the famous Control-F search) doesn't fit into the definition of IR
Van Rijsbergen listed core aspects that differentiate the IR from mere data retrieval: [2]

- IR makes use of probabilistic models
- IR focuses on finding the best matches (fuzzy and also partial answers)
- The user formulates their queries using natural language that not always are even complete
- IR considers only relevant matches

Thus, information retrieval can be seen as the scientific base of what people do in the day-to-day in their interaction with media and other people: filtering the relevant information from the overall information flow. We will focus on the text-based IR because this is the most relevant information for the Content Management Systems (CMS) –that mainly take up the management of text documents.

A. *From data to structured information*

Separating relevant for superfluous information differentiates the portion of data porting significance –that is, information- from the non relevant ones.

The disciplines of IR related to this context are automatic analysis, categorization and classification. Therewith software systems are put in place to analyze the available data to extract information, which is assigned to categories and whose content is classified.

For this purpose, there are very simple procedure that reduce the texts to relevant elements that are normalized, indexed and classified in topics with the help of a Thesaurus, to very complex procedures that relying on the probabilistic information theory (Bayes, etc) are able to retrieve non-sharp results and even more, are capable of learning from the previous findings to retrieve even better result sets.

Using such procedures it is possible for example to assign data to predefined topics according to the information contained, and this for thousand of documents, where a manually procedure wouldn't be possible.

There have been several attempts of standardizing the categories and topics, but nowadays the so called Topic Maps [3] are gaining attention. A topic map can represent

information using topics (representing any concept, from people, countries, and organizations to software modules, individual files, and events), associations (which represent the relationships between them), and occurrences (which represent relationships between topics and information resources relevant to them). They are thus similar to semantic networks and both concept and mind maps in many respects. In loose usage all those concepts are often used synonymously, though only topic maps are standardized.

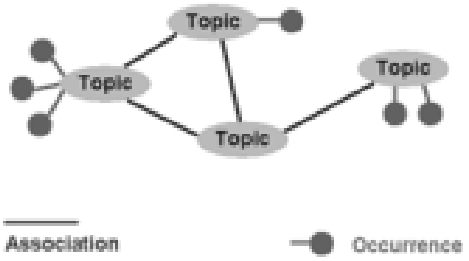

Figure 2

Topics, associations, and occurrences can be typed, but the types must be defined by the creator of the topic maps, and is known as the ontology of the topic map. Ontology can be better defined as "the attempt to formulate an exhaustive and rigorous conceptual schema within a given domain, a typically hierarchical data structure containing all the relevant entities and their relationships and rules (theorems, regulations) within that domain." [4]

To anticipate the applied techniques in Content Management, mentioning that the research in the cognitive information processing area demonstrated that the human brain doesn't manage information in form of tree or hierarchical structures, but nets. [5]

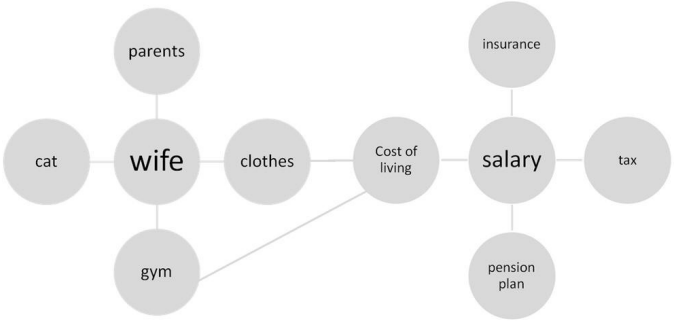

Figure 3

1) Localization of Information

The IR aspects we have mentioned so far take up the automatic analysis, the structuring and can be seen as preparation for the localization, which consists of two procedures: Browsing and search.

Browsing: in the Figure 3 we can see certain relation between the idea of "Gym" and "insurance" going through several information units. The process of following this path of coupled information units where the user's interest can spontaneously change is known as browsing and is of special meaning for Hypertext systems like World Wide Web

Search: the user executes a searching request which semantically represents what the user wants to search.

Search in the context of Information Retrieval differs from search in Data retrieval as following sample shows:
Information retrieval query:

How much can I expect to pay for one hotel during the October beer festival in Munich this year?

Data retrieval query:

Retrieve all documents that contain following words "Price" "Hotel" "October beer festival" "this year"

Especially the word "this year" is critical to establish the difference between IR and DR: in DR, all documents containing the words "this year" are retrieved, whereas in IR "this year" shall be only referenced to the year subsequent to the one of the query.

To sum up, the localization is in IR a semantic oriented process. The problem is that the state-of-the-art of the IR is not advance enough in areas like automatic language analysis, and the query *How much can I expect to pay for one hotel during the October beer festival in Munich this year?* is only processable by means of classification.

III. Information Retrieval and Content Management

We will research which methods are suitable to provide the information consumer with localization capabilities in Content Management Systems. We will provide more detailed definition of Browsing and Search

1) Search: search engines

Google or Yahoo are leading the world wide web searching market. Software products based on them are not only operated by the service providers, but can also be combined with other system and even integrated in the same website. That leads CMS to offer advance searching modules

To provide an idea about the state-of-the-art technologies we will have a look at Google. This search engine combines the search for keyword index available since 2003 -also the indexing of words roots or stemming- with an algorithm to determine the relevance of the documents. This algorithm rates the relevance of one document based on the references to it and counting also the documents referencing these references. Even more, the relevance of the term in the document is determined according to its position in the text or how it is highlighted according to the HTML attributes.
From the semantic perspective of IR, even if Google is a highly developed and highly extended search engine, it is only a keyword indexing based.

The newly created Vivisimo [6] search engine makes a step toward the semantic grouping of concepts and returns the result set like a tree of categories. But even if this engine promises an automatic classification, the tree-view presentation doesn't support optimal browsing.

2) Browsing: Navigation, Sitemap

To supply the desired localization, the Content Management systems provide a variety of Hyperlink list, generated manually or automatically. One very good example of that is the sitemap, which is a structured and coupled representation of the entire website, or the alphabetical index.

Documents in CMS are mostly organized in a tree structured, which also represent the first guidance for the user navigation. Documents are placed by producers at the right position in the hierarchy and are found by the consumers in their navigation at this particular position.

As we already mentioned, the tree-like information organization has been proven not to be sufficient. In the case of browsing, as soon as a document belongs to more than one category, the tree structure doesn't support it and workarounds have to be introduced (like linking, etc)
Another major disadvantage of the tree structures is the lack of target group orientation. Information can be presented in a tree structure in many ways depending on the target audience. Using a single navigation tree leads to sub optimal results only.

3) Looking at the future: the semantic web

Since 2001is the WWW creator Tim Berners Lee and his group tries to answer the question how the information available at WWW can be structured in such a way that not only human readers, but also information systems can process and understand it [7] It relates with the IR in aspects like the automatic semantic processing.

From the semantic perspective the main drawback of the existent WWW documents is the usage of HTML and their focus on the presentation of the information, rather than on the content. As language analysis methods are not that developed to support the IR in documents of unstructured data, Berners Lee's group decided to bring some structure to the documents so that they can be somehow semantically characterized.

For that 2 new languages have been created: RDF [9] and OWL [8] that enable the Ontology description and the tracking of information units and their binding to ontologies (RDF).

A very extensive description on how to create documents for the semantic web is provided by Golbeck [10]. A look at the site "Mindswap", completely based on RDF and OWL is also worth it (*"the first site on semantic web"*)

4) Requirements for the upcoming CMS

The already presented procedures should enable the user to get to their desired information. According to what has been mentioned, there is still a need for the data preparation (manual or automatic), but even for that there are already several well-researched methods.

We will formulate a list of core requirements for the CMS of the near future:

Turn away from the tree-structure: the tree-like representation of the information is no longer support by the state-of-the-art in cognitive sciences. CMS should represent the information by means of topic maps

Usage of IR methods: to build up the information to be managed by the CMS, the IR offers valuable classification methods. These should be analyzed and used in the CMS context, especially to reduce the manual effort of document administration.

Target audience orientation: CMS should be enhanced to offer an access to the information according to the target audience that is at a time requesting it.

Search as standard: the way current CMS administer the complete content of a web site is not optimized for third party searching software or on-the-market search engines. CMS should take into consideration the search engine enabled search

Semantic-Web readiness: due to the big efforts the research community is putting in the semantic web topics, upcoming CMS should be semantic web-ready, which roughly said means a paradigm shift from layout-oriented to semantic-oriented systems.

IV. Conclusion and Outlook Statement

The advantages of a semantic web-ready CMS depending on the user role are from the backend perspective [11]:

- Platform independent content syndication between different systems.
- Access to decentralized data.
- Reduction of the time-to-publish by means of semi-automatic integration.
- Simplified content administration by means of meta-models.
- Content can be compiled

From the front-end perspective:

- Documents are easier to find and the overall search time goes down
- Presentation of subjects for different target
- Description of complex dependencies by means of information context.
- Search optimization by means of nesting possibilities.
- Accessibility to information by means of standardized meta-models

To bring it into the praxis, still a lot of research is required and as starting point can be taken the state-of-the-art of the IR. A very good example of research path would be the application of the IR classification methods to the content data in the CMS.
Another research areas focus on the implementation of semantic nets and target audience orientation.

Even the traditional IR disciplines support the information localization and its administration, but hardly any CMS commercial product makes use of them.

As famous last words, just saying that the CMSs need to make the step towards the semantic web, and to exploit its possibilities the already consolidated Information Retrieval techniques are the perfect tool

REFERENCES

[1] International Organization for Standardization 1974: http://www.iso.ch/iso/en/CatalogueWithdrawnDetailPage.CatalogueWithdrawnDetail?CSNUMBER=7227, 2004-05-12.

[2] van Rijsbergen, C. J. 1975: Information Retrieval - Introduction.http://www.dcs.gla.ac.uk/Keith/pdf/Chapter1.pdf, 2004-04-29.

[3] International Organization for Standardization 2003: ISO/IEC 13250:2003 Topic Maps. http://www.y12.doe.gov/sgml/sc34/document/0322_files/iso13250-2nd-ed-v2.pdf, 2004-06-05.

[4] Wikipedia: Ontology (Computer Science) http://en.wikipedia.org/wiki/Ontology_%28computer_science%29, 2004-06-27.

[5] Schmidt, S. 2004: Semantic Memory. Lecture on Cognitive Psychology. http://www.mtsu.edu/~sschmidt/Cognitive/SemanticMemory.pdf, 2004-06-27.

[6] Available at http://www.vivisimo.com

[7] Berners-Lee, T., Hendler, J., Lassila, O. 2001: The Semantic Web. http://www.sciam.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21,2004-06-27.

[8] OWL Web Ontology Language Reference. http://www.w3.org/TR/owl-ref/, 2004-07-05.

[9] Resource Description Framework (RDF). http://www.w3.org/RDF/. 2004-07-05.

[10] Golbeck, J., Grove, M., Parsia, B., Kalyanpur, A., Hendler, J. 2002: New tools for the semantic web. In: Proceedings of 13th International Conference on Knowledge Engineering and Knowledge Management EKAW02, Siguenza, Spain, Oct 2002.

[11] Koller, Andreas. Content Management Systeme : Ohne Struktur kein semantisches Web. Available at http://www.contentmanager.de/magazin/artikel_1167_wcm_web_content_management_system_semantic_web.html