MANUAL AND AUTOMATED CONTENT INTERLINKING IN A DYNAMIC WEBPAGE

Gabriel Eugen GARAIS, PhD Candidate
Faculty of Computer Science for Business Management,
Romanian – American University, Bucharest, Romania

ABSTRACT

In this article different ways of linking web pages are described using manual and automated content balancing procedures. SEO techniques are also used for better page interlinking with added value given by the new web standard 3.0 that stands on web semantics.
Keywords: web semantics, S.E.O., interlinking, Web3.0,

1. DEFINITION AND NECESSITY OF INTERLINKING

Web Page interlinking stands for the way in which a page is a node of a tree built of web pages and the links between these, the branches having close and narrow design and meaning.

Interlinking is the answer for necessities of:

- Web page optimization
- Reducing redundancies
- Information classification
- Simple imitation of running through different texts
- Pointing to online resource specifications

Interlinking development as a result of optimization necessity and reducing web content redundancy has the ability of increasing the web page readability and correlation power of the content for the visitor.

Structuring the information contained in a web page through classification on categories and subcategories which can have a high hieratical level, has as result a great accessibility to saved data with the possibility of accessing these from different web zones. Web content classification is determined by the heterogeneous nature of data which can be sorted and filtered with easy to satisfy the need of knowledge and rapid access to searched information by visitor or user of dynamic web pages.

The imitation of "turning over" through different pages of a text is a way of basic web page navigation described with technologies of web management.

Creating algorithms for this basic navigation system, is developed further more with the possibility web page interlinking that has at its roots the web semantics. Through this concept that stands as the primary standard for Web3.0, web page connections are created with the semantic meaning of certain words contained in a text at a moment in time. This words can appear in a automated way in a text or by manual define of the person that writes the text.

2. WAYS OF INTERLINKING

This part of article is about the innovative nature by using standards of "Semantic web" understanding. Web semantics offer a common framework which makes possible of sharing, reuse and unify the boundary between applications, companies and community. It is a collaborative work guided by W3C with many scientists and industrial partners, based on the development framework RDF (Resource Description Framework).

The RDF specifications belong to W3C, which was initially developed as a Meta model and then used as a generally model to change the information in different syntax formats. The RDF model is based on the concept of generating information of web resources through the subject-predicate-object triplet, known on the RDF terminology. The subject denotes the resource, the predicate shows aspects of the subject through the expression of the subject and object relation. A way of presenting the model of the RDF triplet in the clause: “Curtea Veche Publishing House sells books Online.” is through defining the subject as “Curtea Vehce Publishing House”, the predicate as “sells Books”, and the object “Online”.

RDF is an abstract model with a variety of serialization formats, and so the particular way of coding a resource or a triplet varies from a format to another.

The defining mechanism of web resources is an important component of W3C for the web semantics standard, being a high level of WWW Consortium, through which applications can store, share and use information in machine language, distributed over Internet, with the benefit given to users of having access to information with a high level of efficiency and certitude. The simplified models of RDF lead to a frequent use in knowledge based management applications.

The serialization formats used by this model are XML and N3.

- The XML format named also RDF, because of the early use among the W3C specifications. His MIME format (Multipurpose Internet Mail Extensions), application/rdf+xml, is recorded as the standard: „RFC 3870”, which recomends that RDF document types should follow the 2004 specifications.
- N3 (Notation 3), RDF serialization format, built as an alternative to XML, being easier to write and follow. Being developed on a tabular structure, it determines a easier way in rendering the triplets instead of using the XML format. The triplets are stored in a triplet database.

The illustration of the two serialization formats shows the general nature of classification and definition of any web resource. It is given the particular web resource “http://en.wikipedia. org/ wiki/Ion_Creanga”, and not knowing that it is actually a link or an article about Ion Creanga within the Online public encyclopedia “Wikipedia”, it is valid that the information according to which the title of this resource is “Ion Creanga” and is published by “Wikipedia”, information that can be written through two RDF expressions.

The triplet form for this information is:

```
<http://en.wikipedia.org/wiki/Ion_Creanga>
<http://purl.org/dc/elements/1.1/title> "Ion Creanga" .
<http://en.wikipedia.org/wiki/Ion_Creanga                    >
<http://purl.org/dc/elements/1.1/publisher> "Wikipedia" .
```

In RDF/XML format this information is written as follows:

```
<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/Ion_Creanga">
          <dc:title>Ion Creanga</dc:title>
          <dc:publisher>Wikipedia</dc:publisher>
     </rdf:Description>
</rdf:RDF>
```

3. SEO FOR INTERLINKING WITH FOREIGN APPLICATIONS, AND THE IMPLICATIONS

The SEO concept is a set of rules that must be implemented in web page development. Behind this concept is the “key of success” for a better web page classification between the search results of search engines and

further more. In other words, a good optimized and structured content can provide a certain level of credibility and eloquence to a web page.

For a foreign web application, here, a possible search engine, to classify the content from an analyzed website, it must include standard structuring elements.

Carrying out the analyze of a page from the "CurteaVeche Publishing" website (http://www.curteaveche.ro/Profunzimile_uitate_ale_crestinismului_Convorbiri_cu_Karin_Andrea_de_Guise-3-725), which contains informations about the book „Profunzimile uitate ale creştinismului" written by Jean-Yves Leloup, it distinguishes classifying elements like:

* it is specified the special charachter encoding, so that the vrowser can show the charachers as it was intended by the publisher
<meta http-equiv="Content-Type" content="text/html; charset=ISO-8859-2">

* a dedicated page title is specified for a rapid identification among other sets of titles
<title>Profunzimile uitate ale creştinismului - Convorbiri cu Karin Andréa de Guise - Jean-Yves Leloup</title>

* a short description is specified for a better understnding of the title
<meta name="DESCRIPTION" content="Carte de la Editura Curtea Veche Publishing: Profunzimile uitate ale creştinismului - Convorbiri cu Karin Andréa de Guise / Les profondeurs oubliées du christianisme - Entretiens avec Karin Andréa de Guise - Autor: Jean-Yves Leloup - An aparitie:2008 - ISBN: 978-973-669-625-1- Colectia:Ştiinţă şi Religie">

* the keyword ar specified for the eloquence of the content
<meta name="KEYWORDS" content="Profunzimile uitate ale creştinismului - Convorbiri cu Karin Andréa de Guise, Les profondeurs oubliées du christianisme - Entretiens avec Karin Andréa de Guise, Autor, Jean-Yves Leloup, 2008, ISBN, 978-973-669-625-1, Colectia, Ştiinţă şi Religie">

* the author and the owner of the current page is specified
<meta name="author" content="Editura Curtea Veche">
<meta name="copyright" content="Copyright (c) 2008 by Editura Curtea Veche">

* thechnical elements of classification are specified dedicated to the search engines which index the page
<meta name="ROBOTS" content="INDEX,FOLLOW">
<meta name="resource-type" content="document">
<meta http-equiv="expires" content="0">
<meta name="revisit-after" content="1 days">
<meta name="distribution" content="Global">
<meta name="rating" content="General">

* a path is specified to a file that contains a structured list of the content in RSS/XML format, which permits the shareing of certain website elements
<link rel="alternate" href="http://www.curteaveche.ro/pagesetter-index-func-xmllist-tid-3-tpl-rss2-pubcnt-1000.phtml" type="application/rss+xml" title="Editura Curtea Veche Publishing - Librarie online">

Furthermore we add the translation and reformatting of parameters generated by the dynamic web application, which resides in the background of the above mentioned web page, through eloquent elements in form of comprehensive text understandable by visitors and other systems that classify the information

contained in other web pages relying on content searched through indexing procedures. The above mentioned link without the coding of parameter looks like:
http://www.curteaveche.ro/index.php?module=Pagesetter&func=viewpub&tid=3&pid=725.

Coding through different programming procedures, results in translating the link, which is a valid like the above mentioned in form of:
http://www.curteaveche.ro/Profunzimile_uitate_ale_crestinismului_Convorbiri_cu_Karin_Andrea_de_Guise-3-725

We must observe the way by which the elements are presented. So, the variable names were intentional omitted, the only important element remaining, the exact place of each parameter. The title was introduces as parameter by replacing the blank spaces with "_", and signs like "diacritical" being replaced with characters from the Latin alphabet. It remains the last two parameters 3 and 725 representing the exact position of the stored information in the database. The three parameters separated by "-" character are extracted from the straight positions and reallocated in order to reveal the searched page in the following order:

- Title= Profunzimile_uitate_ale_crestinismului_Convorbiri_cu_Karin_Andrea_de_Guise
  (The title variable was inserted so that the eloquence of the current page may be easily recognized and classified by other systems that indexes the current page)
- Tid = 3
- Pid = 725

For a quick search of this page in a hierarchy of other pages, it was introduced in classifying fields which permit filtering after a literary type.
The above mentioned elements are about the internal organization of the web page. After calculating and analyzing the level of eloquence in a hierarchy of applications that can be index type, portals or search engines were we can find many keywords from all the pages that were indexed, the obtained position is proofed.
For the above example the search engine www.google.com can be used as a sample. In order to test the descriptive accuracy of the construction of the web page from the above example it was decided to be used that searching in a way that doesn't emphasis the searching of the written word in a successive matter, as it was in the mentioned title of the book. It was chosen a single keyword that can be found in different shapes from a large diversity of competitor web pages and that is "PROFUNZIMILE". This word was random chosen.
The searching of the above word generated a sum of results translated in 15100 independently web pages. The first result generated by google search engine was the page that has just been analyzed in the above paragraphs.
The analyzed webpage is a part of the www.curteaveche.ro website made entirely by the author of this article.
The classification as a position in the list of results returned after the search, is based among the eloquence level, also on a hierarchical algorithm named PageRank, that is a trade mark of Google Inc. company and whose patent belongs to the Stanford University.
The PageRank algorithm is a probabilistic distribution used to determine the possibility that a person that clicks randomly on links to arrive at one moment at a particular page. This algorithm can be used on any volume of documents. The algorithm needs many iterations through the documents collection for a precise result. The probability is expressed with values between 0 an 1. A 0.5 probability, represents 50% chances for a person who navigates randomly through specific interlinked web pages to arrive finally at a specific web page. The diagram of the specified algorithm is as shown in fig. 1

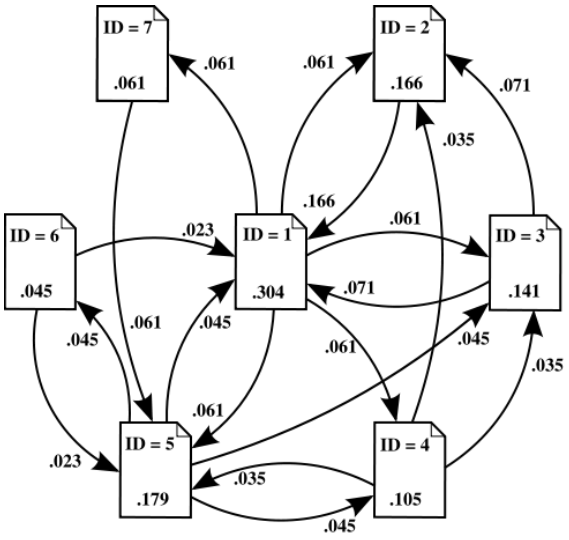

Fig. 1 – PageRank simplified algorithm

The simplified PageRank algorithm:
We assume the existence of 4 WebPages: W, X, Y, Z. The initial approximation of the PageRank is divided equal between pages, so each page start with a value of 0.25.

If X, Y an Z pages link to W pages, each would offer a value of 0.25 to the page W.

$$PR(W) = PR(X) + PR(Y) + PR(Z).$$

where:PR: PageRank results PR = 0.75

Further we assume that page X has a link to page Y and page Z a link to the other three pages. The value of links is divided between all external links of a page. So the page X gives a PR of 0.125 to page W and a PR of 0.125 to page Y. Only a third from the page Z is calculated for PR(W) of aprox. 0.083.

$$PR(W) = \frac{PR(X)}{2} + \frac{PR(Y)}{1} + \frac{PR(Z)}{3}$$

So it came to the conclusion that the PR given to an external link L is equal to the PR of the document itself divided by the normalized number of external links.

$$PR(W) = \frac{PR(X)}{L(X)} + \frac{PR(Y)}{L(Y)} + \frac{PR(Z)}{L(Z)}$$

Through generalization, in this case, PR value for any page "s" is expressed as:

$$PR(s) = \sum_{v \in Bs} \frac{PR(v)}{L(v)}$$

Where:

v: are the pages from which the pages "s" depend from the set Bs (containing all the pages that link to the page s);

L(v): links from v pages.

4. CONCLUSIONS

As it was shown, there are ways of quantification and classification of eloquence level and of the votes number that can a certain web page gather at a specific moment in time through the call of PageRank algorithm, which determines optimization, eloquence, credibility and efficiency in searching a keyword. After implementing the notion of semantic web, in different ways, the information tend to be consistent and relational for an evolutional informational understanding of web page content.

REFERENCES

[1]http://en.wikipedia.org/wiki/Readability_test

[2]http://seattlepi.nwsource.com/
[3]http://en.wikipedia.org/wiki/SMOG
[4]http://www.wordscount.info/hw/smog.jsp (calculator SMOG)
[5]http://en.wikipedia.org/wiki/Flesch- Kincaid_Readability_Test
[6]http://en.wikipedia.org/wiki/Fry_Readability_Formula
[7]http://www.interventioncentral.org/htmdocs/tools/okapi/okapimanual/spachelist.php
[8] http://www.w3.org/2001/sw/
[9] http://en.wikipedia.org/wiki/Page_rank
[10]http://statisticasociala.tripod.com/regresie.htm
[11] Suport baza de date pentru analiza calitativa si cantitativa: http://www.amosnews.ro
[12]http://www.rimeaza.org/definitie-algoritmizare. html
[13] http://www.utgjiu.ro/conf/8th/S2/04.pdf