A Morphosyntactical Complementary Structure for Searching and Browsing

M. D. López De Luise - mlopez74@palermo.edu
Department of Informatics Engineering , Universidad de Palermo University
Av. Córdoba 3501, Capital Federal, C1188AAB, Argentina

Abstract—**This paper is a proposal for the construction of a pseudo-net built with precisely defined tokens describing the content and structure of the original WWW. This construction is derived by morphosyntactical analysis and should be structured with a post-processing mechanism. It is provided also an in-depth analysis of requirements and hypothesis to be stated to accomplish with this goal.**

An in-depth analysis of requirements and hypothesis to be stated to accomplish this goal is also provided. Such derived structure could act as an alternate network theme organization with a compacted version of the original web material. This paper does not describe nor study the post-processing approaches. Instead, it is posted here as a future work. A statistical analysis is presented here with the evaluation of the understanding degree of a hand-made structure built with some tokens derived under the hypothesis presented here. A comparison with the keyword approach is also provided.

Index Term — Web browsing, Web-Mining, morphosyntactical analysis**.**

I Introduction

The special features of the WWW sometimes make it hard, if not almost impossible, to retrieve the exact information searched. Sometimes the results of a search activity include many syntactical matches and even semantic matches that are not related with the actual user searching. There are several well-known solutions for the retrieval activity when searching for specific information. Such solutions normally try to put in context a specific keyword (or a minimum set of them) in different ways: with sophisticated indexing methods, smart visualization alternatives, etc. In this paper provides another point of view: instead of studying the best way to locate and extract specific information, the way to filter and structure some *representative portion* of the content adding the associated site and web location is studied.

This paper presents a proposal (let us name it EC for *Estructura Complementaria* in Spanish, complementary structure in English), a morphosyntactical structure derived from the web that should provide support to certain searches. In this paper it will be shown that it is possible to construct an organized support structure to provide a suitable view of the same data without loosing meaning. This new organization of the data has components with good representation of the underneath information in the top level. Paradoxically, it will be shown here that this morphosyntactical approach can provide help in the construction of an alternate organization with good semantical representation. A set of components, parameters and minimal behavior (i.e. functionality) required to make an implementation of this proposal is presented. The management mechanism of this new construction will not be studied. The detailed structuring algorithm and browsing/querying are not described here.

The remainder of this paper is organized as follows: section II presents some background, related work, the constraints to be met by this proposal and its justification; section III describes the proposal; section IV describes a preliminary test, section V makes a critical analysis of the conceptual strengths and weaknesses of the structure and hypothesis and conclusions; and finally section **VI** states the main work to be done.

II The proposal constraints

As stated earlier, EC is a proposal that depicts some representative portion of the web information and reorganizes it in any other way. It is hard to define the meaning of the word *representative* as it was used in the previous section, but it is the first step to provide a better understanding of any of the requirements stated here.

According to [2] one meaning for the word representative is serving to represent. Therefore, the requirements to be met in order to accomplish the previous definitions are:

I-provide an alternate way to present the same information.
II-provide a compact way to represent more complex and extensive information.
III-provide an alternate structure with the same information.
IV-be the basis for different visualizations of the same information.

In order to work with these requirements and some other derived from the special characteristics of the WWW[1], this proposal has been founded on the following hypothesis:

- The language reflects mental structures that depict a relationship between concepts in a precisely enough way.
- Most part of a language is composed of words or sets of words that represent any kind of objects.
- The contained information in any actual expression of the language can be represented with a number of EC components with a precisely defined alternative structure [14] (this hypothesis satisfy the requirements (II) and (III)).
- There are a finite number of objects in the real world to be represented.
- The searching and mining is user /language dependent (this hypothesis is related to requirement (IV)).
- Each document in the WWW is a kind of conceptual unit.
- There are unlimited resources to process (time and memory).

Furthermore, the following requirements have been stated:
- The EC components must be related to WWW.
- EC must represent implicit information (related with requirement (I))
- A self-adaptive structure.

[1] any of these special characteristics: heterogeneous, very large, dynamical, etc.

In the following subsections a more detailed justification for each one of these hypotheses is provided.

A The language reflects mental templates that depict a relationship between concepts in a precisely enough way

In this proposal the regularity is learned as a structure, which is afterwards clustered automatically without taking into account Chomsky's analysis for generality, precision and completeness. In the actual proposal the derived structures may lack of many, if not all these problems, but this is precisely the way the author handles the divergence between sentence grammaticality and its acceptability.

Noam Chomsky [12] defines the grammar as a product of a machine functioning in an intermediate point of precision between a Markow machine (strongest condition) and a Turing machine (weakest condition). Chomsky labels this as transformational grammar because it is founded in a set of components: set of morphophonemic rules (for phonological components), the phrase-structure and transformational rules (for syntactical components). Each of these components consisted of a set of rules operating upon a certain input to yield a certain output. In his proposal he was centered in syntactical problems and left out the semantic analysis. (O presente o pasado)

Despite the fact that the natural language does not match exactly with the grammar, it has a deep correspondence with it. This hypothesis regards this correspondence to take its regularity as the basis to extract some structures represented by the actual expressions of the natural language.

B Most part of a language is made of words or sets of words that represent any kind of objects

This hypothesis takes the assumption that there is a limited set of objects to be modeled from the real world that handled indirectly within the data.

Chomsky [13] also explains that the language can make infinite use of finite elements (for this infinite use he takes off the traditional concept that the natural order of the thinking is the same as the order in a sentence). These elements are concrete or abstract objects that constitute an internal representation learned from the outer reality.

C An alternate structure of the same information can be generated

This paper takes Chomsky's [12] concept of a subjacent regularity within any grammatical construction and makes a morphosyntactical analysis approach (morphological and syntactical because this proposal is based in words, symbols and the syntaxes related to them) to generate this alternate structure.

The need of alternate structuring and visualizing data can be thought as a consequence of the DB nature of the WWW [21] and makes it possible to help in the searching process [22].

Several proposals [17] have been made to present alternate structures for the web content some of them by metadata manipulation. Some alternatives for such a manipulation are hierarchically structuring of metadata clusters (SenseMaker [15]) for later querying; visually organization of metadata for browsing (Flamenco [16]), Semantic Web construction [18] followed by a Deep Web concept [19], etc.

Lawrence & Giles [1] make a good description of how hard and risky is the task of retrieving information from the Web and the organization and presentation of the web information to a user. This is also true for the semantics approaches. Alan D. [10], Gärdenfors [11], Lesher [20] among others, remark and study some inherent difficulties with the concept of ambiguity, inconsistency and incompleteness that come up with sentences semantically manipulated. Several proposals have been made to manage semantics and its related problems from the Web.

D There are a finite number of objects in the real world to be represented.

This study takes part of this concept: each speaker manages a limited quantum of the objects from reality. This domain of objects is something represented through a limited and numerable set of precisely defined internal elements.

Carlos Peregrín Otero explains [35] that the sentences of a native speaker are related to his intelligence and history. The individual life is too small to be capable to learn each possible sentence exhaustively. Instead, a genetic program enables the learning of the subjacent regularity. These regularities are applied to a set of limited objects of the surrounding reality thanks to natural brain creativity. This creativity is based on certain apprehended rules: there isn't creativity without regularity.

Chomsky also defines a language as a set of statements each one with a finite length and built from a finite set of elements. This, of course, is applied to both artificial and natural languages.

E The searching and mining is user/language dependent

This proposal does not intend to define a visualization mechanism but it has to set the basis to minimize undesired restrictions to its construction.

There are many languages (Spanish, English, French, Japanese, etc.) to be processed by EC, but they must be translated to a common internal language. Browsing the Web could also be hard and basically depends upon the user. This is especially true if the client is not familiar to web activities. Furthermore, the heterogeneity, extent and complexity of the information saved is not helpful.

Many papers have detected and proposed alternatives to solve these problems. Just to mention a few: map visualization and browsing [7], an encapsulating layer that presents a personalized model of the data [23], a visual interface dynamically built with the navigation information [24], to filter customer preferences activating a software agent that learns from user navigation [25], adaptive design based on user profile filtering within a framework [26], to apply Markow chains on user navigation mining data for navigation prediction [27], to show brief information in an affordable visualization the extracted patterns by clustering over the remaining users' web mining [28], to visualize different SOM elaborations from data [29], [30], etc.

F Each document in the WWW is a kind of conceptual unit.

This hypothesis has been stated in order to reduce the complexity of the information structure in the Web, making it easier to model the information. It can be thought as a discard of the analysis for the relationship among sites, delimiting it to each site at a time.

G There are unlimited resources to process

To keep the focus on the main logical analysis, there will not be any hardware restriction considerations. They will be studied deeply in future works.

H An alternate way to read the same information

This proposal is supposed to support a mix of browsing and querying approaches to avoid the typical problems mentioned: it will browse over a set of possible short answers and to query the result of a statistically high probabilistic browsing activity.

From the traditional focus there are two main approaches: browse within a hierarchical structure or querying for specific information. When browsing is performed to search information over a structure, there is typically a set of representing keywords or visual tokens in the sense of (1) and (2) as in the Scent Trails [3], WebEyeMapper[4], GH_SOM[5], Narcissus [6], BibRelEx[8], etc.

But browsing a structure brings up the problem of how to make such structure and when to refresh the contained information. Some papers have studied the resulting impact on the repeatability and success of the navigation process by a user [7], [9]. Some other research have studied the relevance of the query formulation and interpretation ([1], [10], [11], etc.) in the precision and recall metrics.

I The EC components must be related to WWW.

As this proposal intends to be an alternate and compact formulation of the data saved in the WWW, there must be links that allow reaching the original data from the actual data in EC.

J EC must represent implicit information

According to requirement (I), the EC components must be extracted from the WWW and metadata, in such a way that makes them a real representation of the implicit information.

K A self-adaptive structure

The social evolution of the mankind can be easily shown [12], and therefore the consequent language evolution (as it is said to be dependent from the social and biological evolution).

Although many proposals have studied the adaptation to the web content ([6], [29], etc.) as it is a dynamical storage, they do not explicitly consider the language evolution.

III The structure

The Fig. 1 shows the general disposition of the EC proposal as a whole and its relation with the WWW and the users.

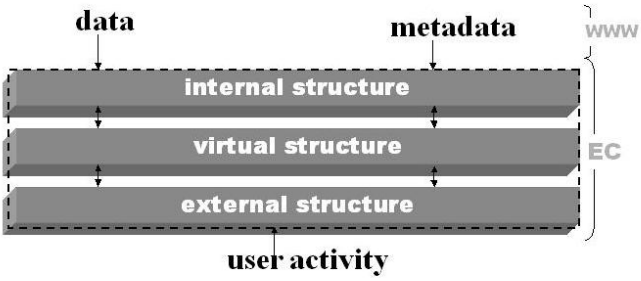


Fig. 1. Three layered structure of the EC. Data and metadata are extracted from the WWW.

The EC uses the data and metadata that are extracted from the actual WWW to feed the internal structure layer wich generate a virtual structure layer. The visual structure layer can be used to get information or browse it.

The three-layered structure of EC is a consequence of the three main activities to be performed:

-*Internal Structure*: gets data from the WWW and process it to derive certain Homogenized Basic Elements streams (HBE). It also solves the gathering, formatting and updating of the raw data and metadata from the Web These steams constitute some kind of translation from the actual language to an internal language. The internal language should be defined to overcome specific language problems. This layer should provide a uniform language vocabulary. It can be thought as an interface between the actual WWW and the Virtual Structure.

-*Virtual Structure*: it analyzes the HBE, detects known regularities or learns new prominent regularities in the stream and organizes these regularities in compact structures (let us name it E_{ci} for the name in Spanish *Estructura de Composición Externa*, External Composition Structure in English, a structure that reflects the closeness of the relationship between words within a phrase). These compact structures are also processed to set a hierarchical structure that point out its corresponding E_{ci} and the original WWW location. This layer should provide a virtual view of the modeled data to the visual structure. This is performed by a set of organized elements labeled here as E_{ce} (for the name in Spanish *Estructura de Composición Externa*, External Composition Structure in English, a supra-structure composed by a set of E_{ci} structures all of them related to the same text. Such structure is intended to establish the way two o more E_{ci}s are related and reflect the main words they are built-up). It is the main part of this proposal.

-*Visual Structure:* to browse or query the virtual structure content as the user information requires it. It should provide alternate models for textual and/or graphical interaction with the user. It should be designed to handle elements like icons, flashing texts, hypertexts, links, etc., to make easier the Information browsing and searching activity. It can be considered as an interface between the Virtual Structure and any user.

This paper is intended to set the main Virtual Structure characteristics, as the other two layers can be thought as interfaces to and from this structure. Thus, it is important to define the Virtual Structure precisely.

A Internal Structure

Basically it processes any data from the WWW taking its specific language as a set or words, numbers and symbols. The specific activity will depend on the language and any language specific activity learned. For instance, one o more words may constitute a HBE, one o more symbols could also be a HBE, etc. (see Fig. 3) The specific behavior should be based on a set of learned rules to let the contents evolve with language. Despite the fact that this is not the central point of this work; a statistical analysis for a reduced set of words will be introduced in a next section as a background to present an example.

B Virtual Structure Description

As stated before, this layer takes a number of HBE streams as input. Then it makes a special processing to generate a set of structures labeled here as E_{ci}.

Some components and their relation are shown in Fig. 2.

MC: (for the name in Spanish *Motor de Composition*, Composition Engine in English, is the E_{ci} -factory) receives from an Internal Structure a set of HBE streams and makes a set of E_{ci}. This set of E_{ci} can be thought as a sub layer within the Virtual Structure layer.

MA (Assimilation Engine) takes a set of E_{ci} from MC and makes a set of E_{ce} structures. Again, this set of E_{ce} can be thought as another sub layer within the Virtual Structure.

Each E_{ci} has derived from an individual statement. Each E_{ce} is composed by one or more E_{ci}. The way E_{ci} are generated depends on the content of the actual HBE stream. The E_{ci} can have one of a number of predefined structures. When MC performs the E_{ci} construction it determines the structure according the presence or not of some special HBEs.

The E_{ce} is generated as a multi-layered structure where the lower layer is created by a sort of association of the E_{ci}-layer elements. The type of structure and the content of the E_{ci} itself determine the nature of this association. The upper-layers are produced as a progressive factoring out of E_{ci} elements.

The actual algorithm for MC and MA can change automatically. The way this is performed will be detailed in the next section.

C Updating process

To update information is not always an easy task in the web. For the metadata derived from this updated data, the consequence is the immediate obsolescence. Updating an EC data should be activated by any change in the original WWW.

In this proposal there is a kind of regulation mechanism for the Virtual Structure layer. This mechanism is implemented by the following main components (Fig. 4):

-A set of Thresholds:

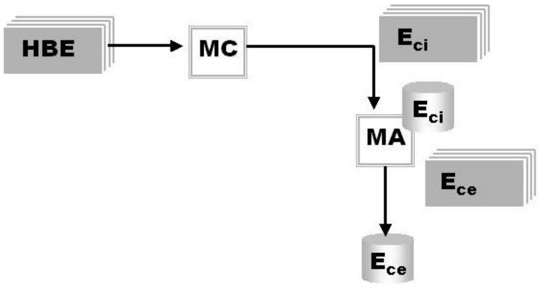


Fig. 2. MC process HBE inputted and generates sets of E_{ci}. Afterwards MA makes a set of E_{ce}.

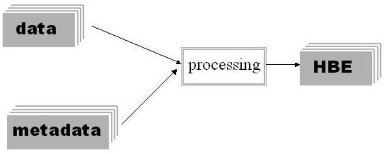


Fig. 3. Translation from any language to an internal language

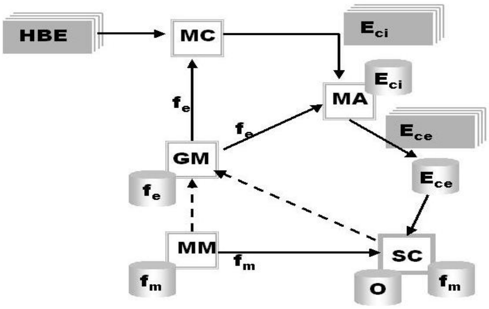


Fig. 4. The updating mechanism.

These thresholds are three and have correspondence with the three levels of processing within EC. The first one is labeled O_{eci} and is related to the E_{ci} sub layer. There is a O_{ece} related to the E_{ce} sub layer and finally there is an O related to the global Virtual Structure layer as a whole.

-A set of metric functions (fm):

The functions *fm* (for the name in Spanish *Función de Medición*, Metric Function in English) state a one-to-one relation with the thresholds O, O_{eci} and O_{ece}. Therefore there are three of such functions: f_m^S, f_m^{Eci}, f_m^{Ece}. They return a numerical magnitude of the relevance of changing the portion of EC that is being revisited. This relevance is some kind of distance between the actual configuration and the corresponding threshold that estimates if EC is the best to reflect the data in WWW.

-A metric functions generator (MM):

As EC works, the actual fm evaluates a sort of distance between the E_{ci} and the actual O_{eci}. The metric function generator can change the way it makes the evaluation. This change will take place when a controller system finds:

$$\mathbf{P(|\ f_m^{Eci}(O_{eci}, E_{ci})| > d)\ ,} \qquad (1)$$

If this probability is high enough, then a new suitable f_m^{Eci} function is generated.

-A set of effect-functions and its generator engine (GM): This set composes the MA and MC functions. Part of the selected subset will constitute the set of temporary abilities for them. These sets of functions are supposed to perform the best suitable activity at that moment. Thus the E_{ci} and E_{ce} processing will change following the actual Threshold values. This set of functions is the core of the activity. The GM engine (for the name in Spanish *Generador de Métricas*, Metric generator Engine in English) makes any change in this set. Such a change could be a deletion, a factoring-out or a modification of any component of the actual set.

For MC: these functions could be learned HBE rules to consider the probability of certain type of structures and rules for reducing, classifying and sorting a stream. For instance, a set can state that if the HBE1 is present in the stream then the structure should be of type 12. But if the HBE500 is also present, then the structure should be of type 5. Let us say the stream is:

HBE2 HBE1 HBE67 HBE20

Let us say also the structure of type 12 requires each of the following HBE following it to be structured as shown in Fig. 5.

For MA: These functions could be a set of learned rules for E_{ce} generations. For instance they could establish that two identical HBEs in different E_{ci}s can be linked as shown in Fig. 6.

-A Controller System (SC):

The SC (for the name in Spanish *Sistema Controlador*, Controller System in English) is a kind of controller that evaluates $P(| f_m^{Eci}(O_{eci}, E_{ci})| > d)$ and fires the GM if it is true. It has also the ability to change one o more thresholds, depending on the range of f_m^{Eci}.

When automatic updating is fired, the specific f_m states if it makes sense to update any portion of EC (it could affect one o more E_{ci}). The updating of one or more E_{ci} cascades up to E_{ce} structures. The same could happen if one o more E_{ce} should be updated.

D Internal Structure as interface

To enable the construction of the E_{ci} and E_{ce}, this layer should solve at least the following problems:

-Translation of text numbers and symbols.

-Avoid translation of less significant words, symbols, etc.

-Consider some kind of special internal word for reflecting the structure of the web as extra information. Cases of such data are: links, hyperlinks, sound, images, etc.

-Consider some kind of processing and later translation of main characteristic for images, sound and video.

-Consider a good translation for ambiguities, contradictions and missing information so as not to miss them as they are considered here as information too.

Finally, note that one o more HBE would represent one o more real world objects. There is no difference in the processing of abstract and concrete objects. The representation problems are solved with the traditional language artifacts.

E Visual Structure

The visual structure should be able to search and/or browse the layered E_{ce} structure for information. The preservation of the references to the actual Web should provide the facility to work on the Web directly if it is needed or desired.

The set of f_m and f_e should select prominent data to be promoted as HBE within upper E_{ce} layers. Less important data should be part of the E_{ci} level.

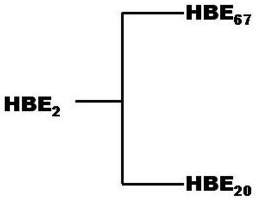


Fig. 5. A visual structuring of a stream.

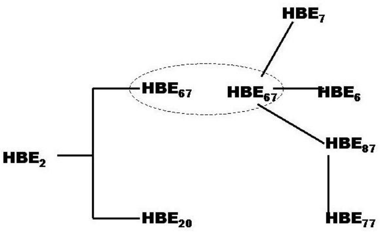


Fig. 6. Two E_{ci}s with identical HBE

As stated before, the visual structure should not have any further restriction if the resulting EC meets the expectations.

IV Preliminary Testing

To find out the possibility of processing E_{ci}s as one of the complementary alternatives for key wording a poll has been made in Buenos Aires (Argentina), in the Universidad de Palermo, and in a private clinic. All the volunteers were native Argentine, Spanish speakers.

A questionnaire with six items had to be answered by 44 volunteers. The items were grouped in two subjects:

a) The four initial questions were related to a decease called *lymphedema*, which has the special characteristic of requiring a long medical treatment. Therefore, patients with this decease usually acquire a pretty good understanding of the lingo and learn some of its related medical information.

Afterwards, the polling was developed with three main groups of volunteers:

-Patients

-Health professionals with deep knowledge of the decease

-Other (mainly informatics students, for better evaluation of part of the test)

An E_{ci} was written instead of a text titled *Normas de prevención de linfedemas* (*lymphedema* prevention). The Original text was never included in the form. Therefore the only extra information was the E_{ci}. It is important to note that the related processing for the E_{ci} is not an optimal one, instead it is the result of a first analysis to approximate this analysis.

b) The two lasting questions were about a topic related to laws. As none of the volunteers had proved to have knowledge of this subject, it was selected to test how good could the inference mechanism be with a few words. These few words were the top level words (the *pointer-words*) from an E_{ci} made with an original lawyer's text.

Following are some of the results obtained.

A Performance of the keyword selecting vs. noun selecting

The volunteers were showed a drawing of one E_{ci} and then asked to write three questions they would like to answer with the original text hidden. They were also asked to write a set of single words to replace each of the questions written before.

To make measurable the relation between writing question efficiency and the key wording, a special processing was done for the questions: the nouns were counted as keywords. In case of ambiguity, the word was also processed as a noun. Afterwards the number of matches between nouns and E_{ci} words were counted. The results were compared with the number of keyword that matched the E_{ci} words.

The Table 1 shows the general data analysis for each set of the noun matches (nm) and keyword matches (km). The Skewness is almost zero showing the typical Normal distribution symmetry around the Mean value. But the negative Kurtosis value denotes a tendency to flat the distribution.

It could be said that the Mean number of matches is higher for nm approach even considering its worse nm value (Mean – Confidence Level) against the best km value (Mean + Confidence Level) with a probability of 0.95.

This tendency is sustained by the Median and Mode values. Conversely, the Standard error, Standard Deviation (and of course the Sample Variance) although comparable, are higher for nm. It should be studied if this deviation could be lowered by improving the algorithmic behind the E_{ci} construction, probably by studying deeply the best set of rules to be applied to words. In the following section some results will be shown, that could be thought as part of this analysis as well.

B Performance of keyword and nouns by knowledge

The same performance comparison as the previous section was performed for each of the three main subsets of the population (see Table 2, Table 3, Table 4).

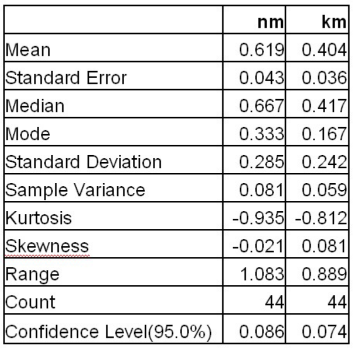
The Mean value for all the sets is always high for nm. This indicates that the average number of matches is higher in all the cases. The same happens with the Median and Mode values.

As a counterpart, the standard error is higher for nm, but the difference narrows as the Count value increases (the number of individuals in the set). This could be the side effect of a small population and should be studied with larger sets to confirm or not the tendency. The Sample Variance (and of course its square root, the Standard Deviation) also presents a behavior similar to the Standard Error.

But in this case the corresponding values in the Table 4 are definitely lower for nm samples.

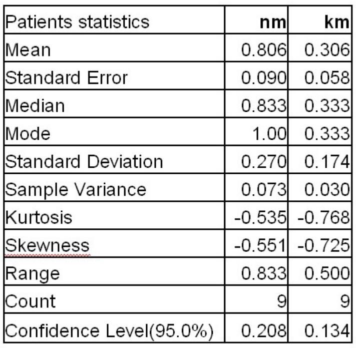
From the Confidence Level (set as 95.0 %) that for the first set worse value (Mean - CL) for nm is higher than the better value for the km (Mean + CL). For the other two sets these values have a small overlap.

TABLE 1
NM VS KM DATA ANALYSIS



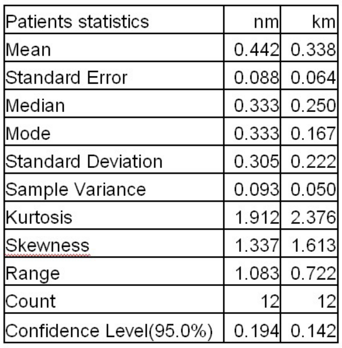
	nm	km
Mean	0.619	0.404
Standard Error	0.043	0.036
Median	0.667	0.417
Mode	0.333	0.167
Standard Deviation	0.285	0.242
Sample Variance	0.081	0.059
Kurtosis	-0.935	-0.812
Skewness	-0.021	0.081
Range	1.083	0.889
Count	44	44
Confidence Level(95.0%)	0.086	0.074

TABLE 2
DATA ANALYSIS FOR THE PATIENT SET



Patients statistics	nm	km
Mean	0.806	0.306
Standard Error	0.090	0.058
Median	0.833	0.333
Mode	1.00	0.333
Standard Deviation	0.270	0.174
Sample Variance	0.073	0.030
Kurtosis	-0.535	-0.768
Skewness	-0.551	-0.725
Range	0.833	0.500
Count	9	9
Confidence Level(95.0%)	0.208	0.134

TABLE 3
DATA ANALYSIS FOR THE PROFESSIONALS SET



Patients statistics	nm	km
Mean	0.442	0.338
Standard Error	0.088	0.064
Median	0.333	0.250
Mode	0.333	0.167
Standard Deviation	0.305	0.222
Sample Variance	0.093	0.050
Kurtosis	1.912	2.376
Skewness	1.337	1.613
Range	1.083	0.722
Count	12	12
Confidence Level(95.0%)	0.194	0.142

Finally, from the Kurtosis and Skewness the population behavior is almost a normal distribution. This is not true for the set in Table 3, perhaps due to the fewer number of samples.

Conclusion: for patients with some knowledge of the topic it has a notable better performance from nm than trying to select a keyword. This is true also for the people with no knowledge in the subject but with less difference with the km. The km sustains its performance in all three sets, but always below the nm alternative.

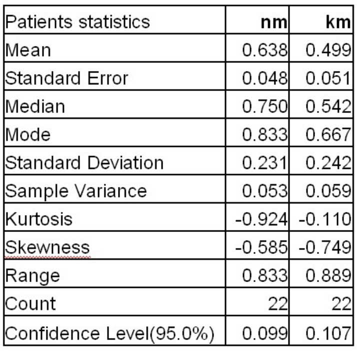
C Performance of keyword and nouns by web searching skills

It was studied from two perspectives: web skills and occupation. The web skill was measured by the quantity of web navigation hours per week. The Table 5 shows the average match for each nm and km alternatives.

Note that each peak in nm corresponds to a decrease in km. Conversely, peaks in km correspond to drops in nm. This behavior should be further studied to determine if it is a markable trend. According to this, there is no evident benefit in having more training in the web. The average value for people spending 31 to 50 hours in the web is almost the same as the value for people spending between 11 to 20 hours in the web. This trend is sustained in the km alternative.

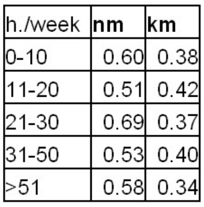
Other external influence when trying to query the web could be the previous knowledge of the user in informatics. It could be thought as other kind of skill. The results apparently confirm this conjecture. Table 6 shows the frequency of nouns and keywords that match with the E_{ci} content. Leyend "S" denotes informatics experience.

TABLE 4
DATA ANALYSIS FOR "THE OTHER" SET



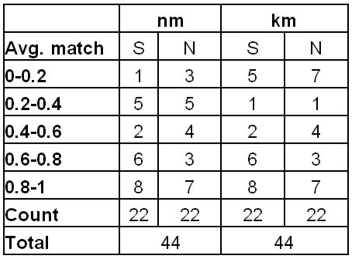
Patients statistics	nm	km
Mean	0.638	0.499
Standard Error	0.048	0.051
Median	0.750	0.542
Mode	0.833	0.667
Standard Deviation	0.231	0.242
Sample Variance	0.053	0.059
Kurtosis	-0.924	-0.110
Skewness	-0.585	-0.749
Range	0.833	0.889
Count	22	22
Confidence Level(95.0%)	0.099	0.107

Table 5
WEB SKILL INCIDENCE



h./week	nm	km
0-10	0.60	0.38
11-20	0.51	0.42
21-30	0.69	0.37
31-50	0.53	0.40
>51	0.58	0.34

Table 6
MATCHING FREQUENCY FOR NM AND KM



	nm		km	
Avg. match	S	N	S	N
0-0.2	1	3	5	7
0.2-0.4	5	5	1	1
0.4-0.6	2	4	2	4
0.6-0.8	6	3	6	3
0.8-1	8	7	8	7
Count	22	22	22	22
Total	44		44	

As can be observed from both alternatives (volunteers with informatics knowledge are half of the total samples), there is a similar number of positive matching but a small improvement is obtained for nm in the low scoring range. These results should be confirmed with a larger sample size.

Conclusion: there is a complementary behavior in the matching performance for the two approaches. Navigation training has no high evident incidence in the matching score.

D Representativeness of E_{ci} information

To evaluate the how the E_{ci} represents the original content, the 44 volunteers were asked to guess a title for a hypothesized text represented by the E_{ci} structure. The Table 7 shows that almost all the guesses were correct. They were also asked to write down three questions that could be answered with the information in the hypothesized text.

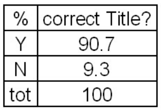
In Table 8 the numbers show a good prediction for the three questions denoted as Q1, Q2, Q3. Note that these questions and title were correct even for those people who did not know about the subject.

Conclusion: the E_{ci} is a good content representation of the original text.

E Representativeness of pointer-words

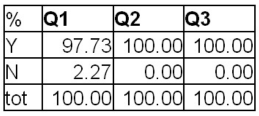
As stated earlier some of the words in the E_{ci} are denoted specially by the rules. It is expected that these words (named here as pointer-words) have a strong influence in the E_{ci} representativeness. These special words had also been tested. A new set of E_{ci}s were developed from a completely new html page.

Table 7
TITLE PREDICTABLY



%	correct Title?
Y	90.7
N	9.3
tot	100

Table 8
TOPIC PREDICTABLY



%	Q1	Q2	Q3
Y	97.73	100.00	100.00
N	2.27	0.00	0.00
tot	100.00	100.00	100.00

To assure null previous knowledge for all the volunteers, the subject was a new law regulation and its social consequences. Six pointer words were deduced. The polling form asked to guess three titles for this text using these words as the only knowledge about it.

Table 9 denotes that even with these few words, represented here as the column labeled with t1, the main title had a good chance of being correctly guessed. It is interesting to see that the first guessing was always the best one, since the other two titles have a lower probability to be correct.

To evaluate the convenience of adding location information, the same question was repeated after giving the domain of the page. Table 10 shows that it could serve as disambiguation information.

Conclusion: pointer words could be useful as text representation. Therefore they could be a kind of reduced index for the E_{ci} . The E_{ci} also could be a kind of index to the original text. These results should be studied for larger samples for a better verification.

V Conclusions

A data treatment approach that relies on the concept that the data can be reduced losing and reordering data was presented. For this treatment a morphosyntactical structure was created. This structure is not designed to handle semantics directly but, the syntax and morphology of the language:

-Eci represents a good representation of keyworking.
-Eci performs well for people indepently of their knowledge on the specific subject
-Eci performs well for people indepently of their informatic knowkedge.
-Navigation training has no high evident incidence in the matching score.
-The Eci is a good content representation of the original text.
-Pointer words could be useful as textreduced index for the Eci.
-The Eci structure could be a kind of representation of the original text.

Remark: Virtual Structure is not a visual structure. Any figure or visual representation in this paper is for better illustration of the proposal.

VI Future work

There is a set of pending topics to be studied detailed:

-An implementation of an automatic adaptation to the dynamic structure and content of the data in the Web.
-The best selection of the words and the rules.
-The internal language, to overcome specific language problems (ambiguities, missing data, contradictions, etc.)
-A good design for the Visual Structure should be defined.
-A suitable treatment for multimedia data. This data should qualify for processing by EC defined components.
-Algorithmic and performance aspects.
-Algorithmic alternatives for generation and refinement of f_m and f_e functions.
-Extension of the HBE translation for special symbols.
-Extension to other languages.
-Performance and evolution of EC components with different languages.

TABLE 9
PERCENTAGE OF CORRECT GUESSES FOR THE TITLE

%	t1	t2	t3
Y	88.37	76.74	69.44
N	11.63	23.26	30.56
tot	100.00	100.00	100.00

TABLE 10
PERCENTAGE OF CORRECT GUESSES FOR THE TITLE AFTER URL INFORMATION

%	t1	t2	t3
Y	100	100	100
N	0	0	0
tot	100	100	100

-Consider avoiding the hypothesis that states: Each document in the WWW is a kind of conceptual unit.

-Consider avoiding the hypothesis: There are unlimited resources to process.

ACKNOWLEDGMENT

The author gratefully acknowledges the contributions of Dr. J. Ale and Prof. M. Bosch for their work on the original version of this document.

REFERENCES

[1] S. Lawrence, C. L. Giles, “Searching the World Wide Web”, Science vol 280. Pp 98-100. April 1998.

[2] Merrian Webster Dictionary. http://www.m-w.com/

[3] C. Olston, Ed H. Chi. “Scent Trails: Integrating Browsing and Searching on the Web”. ACM Trans. on Computer-Human Interaction, vol 10, No. 3, September 2003, pp 1-21

[4] R. Reeder, P. Pirolli, S. Card, “WebEyeMapper and WebLogger: Tools for Analyzing Eye Tracking Data Collected in Web-use Studies”, UIR Technical report UIR-R-2000-06

[5] D. Merkl and A. Rauber, “Uncovering the Hierarchical Structure of Text Archives by Using an Unsupervised Neural Network with Adaptive Architecture”, Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD2000)

[6] A. Wood, R. Beale, N. Drew, B. Hendley, “HyperSpace: A World-Wide Web Visualiser and its Implications for Collaborative Browsing and Software Agents”, HCI'95

[7] H. Chen, A. Houston, L. R.Sewell, B. Schatz,” Internet Browsing and Searching:User Evaluations of Category Map and Concept Space Techniques”, Journal of American Society for Information Science (JASIR).vol 49.nro 7.pp 582-603

[8] A. Brüggemann-Klein, R. Klein, B. Landgraf, “BibRelEx-Exploring Bibliographic Databases by Visualization of Annotated Content-based Relations”, D-Lib Magazine.Vol 5.ISSN 1082-9873

[9] H. Chen, C. Schuffels, R. Orwing, “Internet Categorization and Search: A Self Organizing Approach”, Journal of the visual communication and Image Representation.vol 7.pp 88-102

[10] D. Alan, “A Comparison of Techniques for the Specification of External System Behavior”, Computing practices 1998.

[11] P. Gärdenfors, “Meaning as Conceptual Structures”, Tech Rep LUCS 40.Lund University Cognitive Studies.ISSN 1101-8453.

[12] N. Chomsky, ”Syntactic Structures”, Walter De Gruyter, Inc. 1974. ISBN: 9027933855.

[13] N. Chomsky, “Aspects of the Theory of SyntaxMIT Press. 1969. ISBN 0262530074 9.

[14] N. Chomsky et al., “Langue : Théorie Générative Étendue”, Hermann. 1977. ISBN 2705658394.

[15] M. Baldonado, “An Interactive, Structure-Mediated Approach to Exploring Information in a Heterogenous, Distributed Environment”, Ph.D. Dissertation, Stanford University, December, 1997.

[16] A. Elliott, “Flamenco Image Browser: Using Metadata to Improve Image Search During Architectural Design”, Ame Elliott, Doctoral Consortium, in the Proceedings of the ACM CHI 2001.

[17] E. Amitay, “Web IR & IE”, http://www.webir.org/

[18] T. Berners-Lee, J. Hendler, O. Lassila, “The Semantic Web”, Scientific American. May 17, 2001.

[19] M. Bergman, “The Deep Web: Surfacing Hidden Value”, Bright Planet. July 2001.

[20] W. G. Lesher , J. Moulton Bryan, D. J. Higginbotham, “Effects of Ngram Order and Training Text Size on Word Prediction”, Dep. of Communication Disorders and Sciences. Univ. of New York at Buffalo.

[21] P.A. Bernstein, “Panel: Is Generic Metadata Management feasible?”, Proc/ of the 26th Int. Conf. on Very Large Databases, Cairo, Egypt, 2000

[22] R. Davis, H. Srobe, P. Szolovits, “What Is a Knowledge Representation?”, AAAI. SPRING 1993. Pp 17 – 18

[23] J. C. French, E. K. O'Neil, A. Grimshaw, C. L. Viles, “Personalized Information Environments”, Poster. Darpa Contract N66001-97-C-8542

[24] A.M. Wood, N.S. Drew, R. Beale, R.J. Hendley, “HyperSpace: Web Browsing with Visualisation”, Third International World-Wide Web Conference Poster Proceedings, Darmstadt, Germany, April, pp 21 – 25

[25] A.S. Pannu, K. Sycara, “Learning Text Filtering Preferences”, Proc. of the AAAI Symposium on Machine Learning and Information Access (Stanford,CA,USA). 1996.

[26] M. Perkowitz, O. Etzioni, “Towards Adaptive Web Sites: Conceptual Framework and Case Study”, Dep. of Computer Science and Engineering. Univ of Wachington. Seattle. USA. Artificial Intelligence 118 (2000) pp 245 – 275.

[27] B. Trousse, ‘Evaluation of the Prediction Capability of a User Behaviour Mining Approach for Adaptive Web Sites”, Inria - AID Research Group, B.P. - Sophia Antipolis Cedex. France.

[28] I. Cadez, D. Heckerman, C. Meek, P. Smyth, S. White, “Visualization of Navigation Patterns on a Web Site Using Model Based Clustering”, Dep. of Information and Computer Science. Univ of California, Irvine. 2000.

[29] K. Langus, “Text Mining with the WEBSOM”, Acta Polytechnica Scandinavica. Mathematics and Computer series No 110. 2000.

[30] K. Langus, T. Hokela, S. Kaski, T. Kohonen, “Self-Organizing Maps of Document Collections: A new Approach to Interactive Exploration”, Simoudis E. Han J. Fayyad U. eds. Proc of the second Int Conf. On knowledge discovery and Data Mining, pp 238-243.AAAI Pres. Menlo Park. CA. 1996.

[31] L.G. Heings, D.R. Tauritz, “Adaptive Resonance Theory (ART): An Introduction”, Technical Report 95-35, Leiden University, 1995.

[32] M. Jaczynski, B.Trousse, “Selective Markov Models for Predicting Web-Page Accesses”, University of Minnesota, Department of Vcomputer Science/Army HPC Research Center Minneapolis. 2000.

[33] Britannica online.

[34] S. Kaski, “Data Exploration Using Self Organizing Maps”, Acta Polytechnica Scandinavica,. No 82. Dr Tech. Thesis. Helsinki University of Tech. Finland. 1997.

[35] P. Otero, Introd. In ”Estructuras Sintácticas”, Siglo XXI editores SA. 1974. Siglo XXI editores SA. 1974. ISSN: 1139-8736