Semantic Search

Philosophical approach - From ideas to concepts

Dr. Juan Chamero, jach_spain@yahoo.es, *Darwin Architect, Buenos Aires 10 February 2009*

Something about the deep mechanic of knowledge searching

The paradox of the man as hunter/fisher of Information and Knowledge

When people have information needs start a not well known yet mental process that probably inspects first their own cognitive assets and as a probable outcome "ideas" prompt within their minds, perhaps of similar nature to the ones prompted when being hungry or thirsty, If challenged to explicit those **"information needs activated ideas"** they will try to do it via oral and/or written symbols of a given language. If they were challenged to mark from 0 to 10 how much they "know" about those ideas they will do it as well. Finally no matter how much they confess to know as they are human and as such curious creatures most of them will try to know more and better about their needs in order to be fully "satisfied". However the knowledge as a nutrient behaves strongly cumulative but paradoxically within the same physical storing volume and a significant part of it lasting for long. The scheme would be something like this:.

Something to look for => "idea in mind" prompts => representation (exteriorization) of it in a given code

And as this nutrient has to be fished, retrieved from a sort of World Oracle, fished out of a Web Ocean, the fisher needs of a strategy, adequate tools and weapons and an adequate bait for example a sequence of "words" and/or logical operators belonging to a given Jargon:

("a b", c AND d, NOT "e f"),

meaning to look for documents that have the precise chain aSPACEb in their text corpuses where for example a and b are words of a Computing Jargon in American English and SPACE stands for a "written" space as in:

"parallel computing"

a chain of two words and a single space between them, and in more detail a sequence of 18 characters as follows:

"P, A, R, A, L, L, E, L, SP, C, O, M, P, U, T, I, N, G)

where characters including the SP (Space), could be represented in a particular 8-bits code as a 144 bits chain. Immaterial commas are included here to clarify the meaning visually. Some search engines like Google use these types of words chains interpretation when words are within "quotation marks" as we put above "a b" ⇔ "parallel computing".

Usually people make use of a sequence of words evenly separated by spaces, generally one, as for example: **parallel computing** without being bracketed by "quotation marks" or some other equivalent character In this case most search engines look for documents that somewhere within their text corpuses have the words **parallel** AND/OR **computing**. Some search engines control the words distance between them in order to consider the "match" as valid and some others do not. Some consider text corpuses as single vectors not interrupted by punctuation marks meanwhile some others do not.

Actually in the Web Ocean only words are indexed, no concepts

The whole expression above ("a b", c AND d, NOT "e f") could then be considered a **"query"** to a **"Knowledge Database"**, usually structured and "seen" by conventional Search Engines as a huge virtual two-dimensional array of "documents" – "words", namely 20,000 million documents versus a few million of distinguishable" words" per language.

Among distinguishable words we may find **"Common Words and Expressions"** ([1]), and a myriad of ***single word*** acronyms, neologisms, bad written but acceptable as valid common words and expressions, toponymics , names of persons either physical or juridical, tools, mechanisms, methodologies, procedures, etc. However as we are going to see the most numerous set of concepts is missing.

Effectively most actual search engines do not index concepts such as "parallel processing", ""quality of education", "big-bang theory", "attention deficit hyperactivity disorder", and even well known names as "Albert Einstein" and "Isaac Newton". What are usually indexed are single words "equivalences" as QOE (with at least two main acceptations: Quality of Education and Quality of Experience), ADHD by the attention disorder, possible names as Einstein and Newton but encompassing all possible single and multiple words homonyms and acceptations. This missing is crucial because it impedes the semantic search. But even though actual search engines were able to detect and index documents by all meaningful single and multiple words **"keywords"** the semantic ambiguity still would persist!. Why?: because the same keyword, no matter if single or multiple word may have multiple meanings as a function of subjects deal with. The single word keyword "pond" for instance may be meaningfully used in more than a hundred of different knowledge matters.

In brief conventional search engines work with huge but very limited virtual arrays, from the point of view of semantic, because they index documents only by single words and some concatenation of words seen as single words like for instance "big-bang"-

How concepts are unveiled

Darwin, a distributed intelligent agents methodology of **"Knowledge Discovery"** intents to go farther, to unveil from existent Web documents all meaningful word chains and attach to them their **"semantic path"**, the meaning domain where they belong transforming them, by "de facto", in **"concepts"** instead of remaining as potential and ambiguous keywords. This rather heavy Knowledge Discovery task implies a previous documents semantic classification in order to find for each document its **"main subject"**!. These subjects are somehow hierarchically ordered along very specific "semantic paths" as well.

Common Words and Expressions are no more than a few hundred thousand but combining them wisely experts of the different knowledge disciplines were creating along the time millions of concepts per language. These concepts are distinguishable within documents written, by and for, intelligent beings. And one of Darwin conjectures is that other intelligent beings and specially settled and trained agents may unveil them.

Human beings as Knowledge Databases users behave as almost inscrutable "black boxes". Seen from databases side users behave like black boxes that issue queries supposedly oriented to satisfy their "information needs". Seen them individually it is almost impossible to guess or to infer with a reasonable probability of success what they are looking for. The above described **"ideas in mind"** are extremely variable and fuzzy along the time even for the same person and for the same stimulus and also extremely variable and fuzzy their associated queries.

What´s in an idea

To understand better the subtle differences among images, ideas, words, ideograms, keywords, concepts, meanings, and subjects we think it is necessary to deep a little in epistemology ([2]) a branch of philosophy that study the nature of knowledge trying to answer the following questions: what´s knowledge?; how knowledge is acquired?; what do people know?; how do we know what we know?. It seems that what we know about "information" is still limited to ***Claude Shannon*** **"Theory of Information"**. From those times (1948), we humans are almost in the same place except that our capacity to process information in the computing domain of [memory – speed of process] have stepped up billion times. However we still ignore what knowledge and intelligence are in despite of spectacular advances in AI, Artificial Intelligence. We are absolutely convinced that humans have the most advanced intelligence in the universe at our reach and at the same time we humans have created forms of artificial intelligence that challenges ours and in some instances beat us. We are pragmatic accepting that we are machines of thinking and that perhaps thinking be a new advanced sense and that perhaps the intelligence be a fluid more subtle than information or perhaps a quality to see the immanent and most times hidden order of the universe. Could the scientific challenge in this area be imagined as the sequence **material mass => energy => information => intelligence?.**

Something about intelligence that lay "behind" documents

We and many colleagues use and argue about the concepts of "unveiling intelligence", "knowledge discovery", "retrieving intelligence" and many audacious expressions of this sort. What we mean by that?. Let´s go back to see critically the best actual search engines performances. They intent to offer us universal indexes of almost everything humans have documented, word by word, like having a celestial map of the whole universe, particle by particle. Is that enough?. Unfortunately no!. Intuitively we dare to say that something like "intelligence" is missing, don't we?. Of course we are sure that every document has its proper intelligence, the one that was wisely architected by its authors. If the corpus text of each document were written "mathematically" we may argue that perhaps this "hidden" intelligence is closely related to a WFF, Well Formed Formulae that concatenates words and symbols. But unfortunately these corpuses are literary, written under complex but rather fuzzy rules in order to be these "messages" understood by other humans within a wide range of comprehension.

Ten years ago we devised a set of **Darwin Conjectures** about how humans document, primitive rules of thumb that if found statistically true will enable us to envisage the cognitive core of the document, something like a meaningful abstract of it. This task resembles the old and patient task of librarians doing the card index for each book. **Darwin Methodology** does the same but performed almost autonomously by agents. Darwin primitive "intelligence" retrieved has the following form:

> The knowledge domain (discipline) to which the document belongs;
> The main subject dealt with;
> The semantic path, from the discipline root till the main subject dealt with:
> The "**keywords profile**", that is the literary concordance of keywords retrieved versus the main subject keywords set;
> Abstract of the document, something like its "**document fingerprint**", a statistics weighted vision of the remaining corpus obtained by eliminating non-keywords;

This primitive intelligence could be progressively enhanced via a fast and convergent learning process. Let´s see now what could we learn from philosophers and great thinkers.

The Idea ([3])

As per Plato

For Plato "real" things are hosted in the realms of "forms" and "ideas". No matter here to discuss if they are either finite or infinite in numbers or if they are either forever existent or cumulative along time. What´s important for our reflections is the coherence of Plato´s ontology suggesting that images prompt in our minds pointing to "pre-existent" ideas most of them "old ones" and by exception, from time to time, "new ones". These ideas rest within our space-time reality but are at large preexistent in an upper level conscious realm. Perhaps in Plato´ terminology idea was similar to what we name as "concept".

As per Descartes

For Descartes ideas were images but not necessarily existent "in mind". He rests more on the ground that Plato saying that not all our thoughts are images of "things" and only those images deserve the name of ideas sustaining that they are "innate", in practice close to the Plato vision of a pre-existent realm. By the way in Zen practice students work trying to think both ways, with and without images!.

As per John Locke

For John Locke idea is whatever behaves as the outcome of thinking, something like saying: If I´m thinking so I´m developing ideas. He dared to qualify good thinking as "good sense", experimenting outcomes "down to earth".

As per Hume

For Hume the idea is the outcome of a process of thinking about perceptions. Pragmatic but not bad for our practical purposes!. We know what is outside us, via life experiences and at large via perceptions, ours or derived from others.

As per Kant

Please let stop a little here because for Kant idea opposes to concept meanwhile for our ontology concepts are derived from ideas. Let's try first to understand what Kant meant by that opposition. A man may have an idea about something, putting an example within our ontology "parallel processing" within computing and next within programming. John has an idea about it and deepening enough he dared to define it as a "concept". For him his definition is a concept because following Kant it refers to a very specific definition, that in his criteria it should be accepted universally. But this is only John´s opinion. And we may imagine that for the "same" phenomenon there probably appear hundreds of different opinions. Kant talked about "regulator ideas" or ideals that people (we, not Kahn, dare to say statistically) tend to follow. Then without contradicting Kant we may argue that people thru thinking, thru generating ideas, and statistically moving around "hidden" regulator ideas may eventually create "concepts" that have a restricted and limited life but sound enough to create knowledge, let´s say the knowledge of a given civilization at a given time. These concepts could be considered as "modals", "dominants", acceptable as the best definitions for a knowledge realm, but impermanent, however enabling us to "map" the knowledge "as_it_is" at a given moment. This is very important because what Darwin agents retrieve for us when mapping the Web as_it_is are in fact **modal concepts**!.

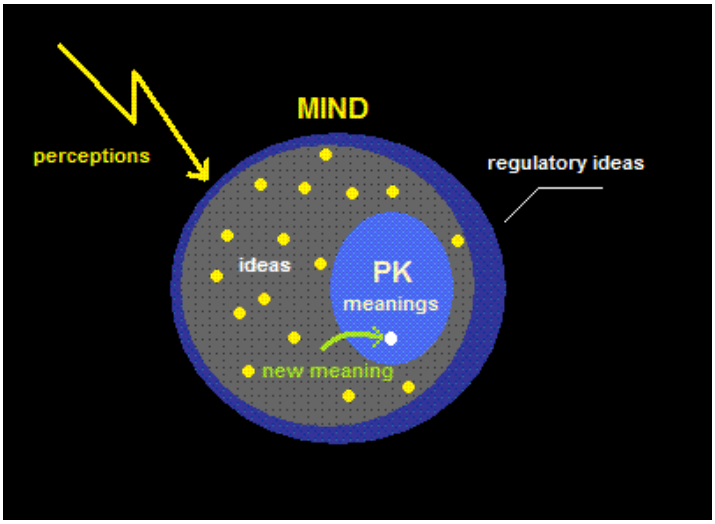
As per Steiner

Rudolf Steiner in the line of Goethe's thinking launches a very interesting idea saying that perhaps thinking is the outcome of a new organ like the eye, to "see" reality with a new perception. As the eye perceives certain light wavelengths and ear sounds of certain wavelengths also the "thinking organ" perceives ideas. In our Darwin ontology we assimilate concepts as **"wavelets"**, behaving intuitively like semantic particles that are born of a sudden within a given discipline, have a certain "lifetime" and finally obsolesce and die.

As per Spinoza

We shall not enter here in the Spinoza' idea of ideas because its complexity. Let´s keep as valuable for our analysis his idea of types of ideas: true, fictitious, dubious, and false. The true idea, is unattainable, beyond our reach, however all ideas derive from it. Fictitious ideas are born from a fiction, are ideas that we "make believe" their existence or inexistence. False ideas are derived from fictitious and are due to errors in our reasoning. Finally dubious are ideas that we "see" as fuzzy, far from clear and distinctly. For our purposes we work with fictitious and dubious ideas. We believe that in the Web space authoritative documents host fictitious ideas meanwhile queries from people interacting versus search engines databases host dubious ideas.

Something of Imagery: from ideas to meanings



The figure above depicts a free interpretation of Kant idea of "ideas" as "seen" from our Darwin K side Ontology (see below K Realm). Perceptions trigger in our mind ideas via a process not known yet. PK, Personal Knowledge would be the Personal asset of "meanings". Kant talks

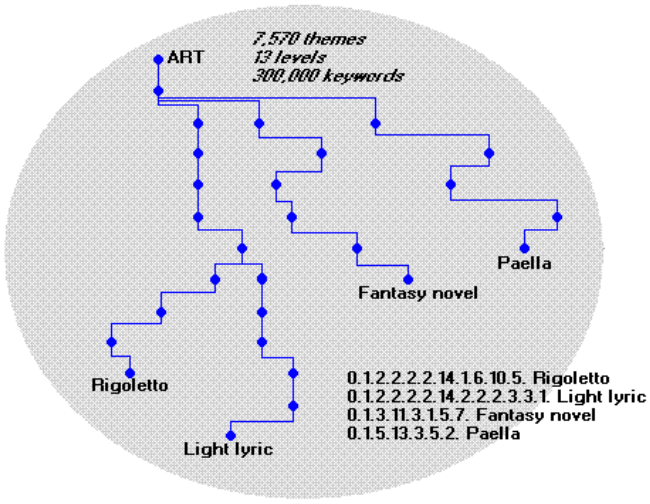
about “regulatory ideas” that exist somewhere and that for him are innate ([4]). Then some unknown yet intricate process involving the interaction of perceptions, PK, plus Regulatory Ideas prompts ideas in our mind. From time to time new meanings may appear that are somehow hierarchically located within PK being among new meanings those that update and/or transforms PK, some ones in large extent. For example meanings that involves new visions.

Concept

Finally we arrive to “concept”, for many authorities the basic unit of knowledge.
I’m inclined to see concepts as units of meaning instead. In Darwin ontology a concept correlates to a definition located “hierarchically within” a given knowledge domain. And “hierarchically within” implies that it is located at the end of a unique “semantic path” within a knowledge domain. The word “unit” and “basic unit” could be misleading because it induces in our mind members of a set sharing the same or similar hierarchy.

As in Darwin vision formal knowledge structures itself over inverted **“logical trees”** all paths could be considered units within a hierarchically structured topology. These paths have two extremes: the “head” pointing upwards the tree and the “tail” pointing downwards the tree. If we imagine a certain Human Knowledge domain structured like a tree the top head is the “root“ usually represented upwards on top meanwhile derived subjects are represented downstream to the most specific subjects: the “leaves”

Note 1: Common Expressions are long coined concepts that universally prompt in our minds well defined situations, scenarios, conclusions, wisdom flashes: Latin locutions like “Res non verba”, “sine qua non”, proverbs, citations and sayings, like “first comes first served”, “a burnt child dreads fire”, “a one thousand journey starts with the first step”, and even short quotations like “we burn daylight” (Shakespeare).
Note 2: epistemology, from Wikipedia
Note 3: idea, a discussion about idea and ideals from Wikipedia; Baruch Spinoza, from Wikipedia
Note 4: Take care; the “innate” of Kant is not the same as the “innate” for Plato .Humans have mind restrictions and “condemned” to see reality thru tinted glasses.



The figure above show us several “semantic paths” of ART as_it_is in the Web as per August 2008, a knowledge domain unveiled by Darwin agents: 7,570 themes along thirteen levels holding up to 300,000 keywords. “Rigoletto” is a single word keyword hosted at the “end” of path [0.1.2.2.2.2.14.1.6.10.5] as its “tail” being 0 its “head”, the ART root.

Creation of a new concept hypothesis

When we humans “create” a new concept it is agreed that we have arrived to a “precise enough” definition of an “ideal”, following a mental collective process of thinking in the neighborhood of Kant “regulatory ideas”. This definition has sense only if referred to its precise context within the semantic space of the knowledge. So if something like the fake “***Programmers Collective Authority***” agrees about the meaning of “parallel processing”, this definition has sense along a path of the form

[Information Technologies and Communications => Information Technologies => Computing => Software => Programming => Programming Languages …..]

That is a semantic path with its head on the IT&C root and its tail pointing specifically to a subject such as for example “parallel processing” in Operating Systems. Tails end in tree nodes, including the roots. And it is highly probable that he same name: “parallel processing”, will be used by other humans to define agreed meanings for other semantic paths, for example in chemistry, economy, drugs industry, etc.

Keywords

Now let´s face the meaning of this rather confuse concept. In the Search Engines industry are words or chains of words that “magically” open Pandora boxes that hide pieces of knowledge we, humans, are looking for, bringing documents references that “prima facie” satisfy our cognitive needs.

We have to take into account that Darwin technology working in the Web space deals with two realms: the “**K Realm of the “Established Knowledge**” represented by all Web sites and the “**K´ Realm of People”** navigating by the “Web Ocean”. In fact Darwin Technology works based in two interacting ontologies one for each realm.

Real existent Web keywords are in fact “created” by “authorities” initially most of them as neologisms, new words formed by combining in a precise way pre-existent single words sequences like for instance the rather old **outlet** that probably was initially imagined as out-let, **up-stream**, **down-stream**, **well formed formulae** that ended as **WFF**, and some other of recent creation such as **"quality of education assessment"**. In fact these keywords could be considered “new words”, new ideograms created to facilitate the human communication and to make it more universal and precise. Perhaps they should be written concatenating their components using (_) or Upper Case letters like for instance **WellFormedFormulae** as equivalent to **well_formed_formulae** or **AsItIs** equivalent to as_it_is.

Keywords versus Common Words domains

Keywords domains are very limited, restricted to a specific subject within a given discipline meanwhile Common Words and Expressions domain is the one of the full language to where they belong. For instance the common words “well”, “formed” and “formulae” are valid for the English language no matter the subject deal with. On the contrary the keyword name “parallel processing” has existence in at least 100 domains –subjects- as tails of its 100 associated semantic paths.

In languages like for example Chinese keywords are represented as new ideograms because is what they are: representation of new specific ideas. Along time some keywords become Common Words or Common Expressions and many finally die, usually by obsolescence. The same but slowly happen to Common Words and Expressions.

The semantic space of existent documents (within the Web space)

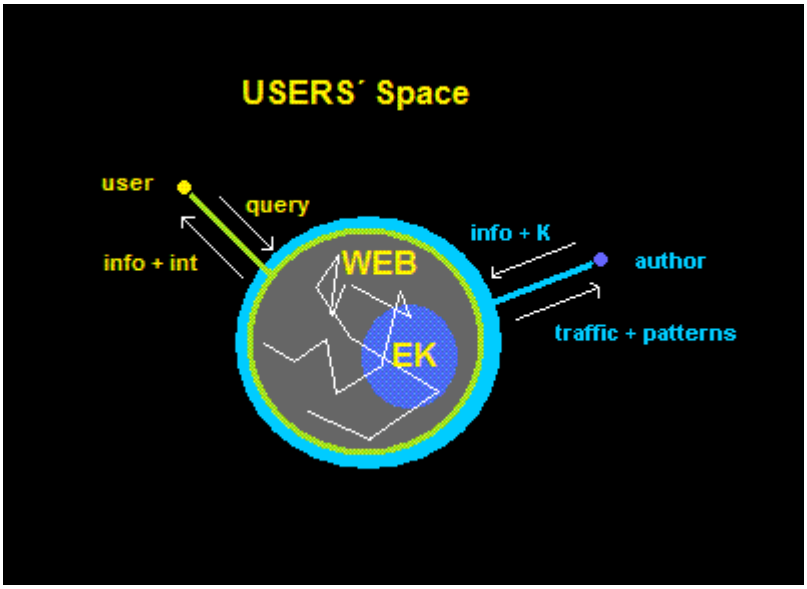
Primitive indexes of this space are located in main Search Engines databases basically structured as virtual two dimensional arrays of documents versus words. Actual search engines do not classify by keywords, only by accepted “words”. These arrays are huge, in the order of 20 million “columns” one for each document hosted and one million “rows” one for each common word or expression, including brand names, geographical names, personalities names, and well known acronyms.

However something is missing: concepts. As they are not detected and if declared by documents´ authors are either neglected or considered not credible they should be “unveiled”. Darwin ontology of K side guide Darwin agents to perform this important task. Let´s suppose that we were able to unveil the main subject of each document –whether unique- and reorder the documents-words arrays –at least one for language- putting together documents that share the same/similar main subject. By studying these document clusters from the point of view of literary concordance we are going to detect combination of words that tend to appear regularly in most documents of each cluster, and that at the same time are “rare” enough within the whole Web universe to be considered a Common Word or a Common Expression, and that persistently tend to appear associated to others belonging to the same set of rare combinations.

This characteristic means that the same combination could exist in some other clusters, belonging to different disciplines and even to the same discipline but never associated to the same semantic neighborhood. For example the combination “parallel processing” related to Computing may appear in another cluster of the same discipline, for instance having as neighbors “n-tier”, “multitasking”, “interleaving”, “distributed computing”, but it may also appear related to Human Brain Processing, associated to “neural network”, “incoming stimuli”, and

“computer vision”. Of course their definitions: parallel processing in Computing, and parallel processing in Human Brain Processing are completely different and surely their documented definitions are somewhere in more than one Website but they are not easy to find in this early stage of Knowledge Discovery. One of the Darwin K-side conjecture says that parallel processing located at the tail of a path of the Human Brain tree should be a different concept of parallel processing located as the tail of a path of Computing. At the actual state of the art of Knowledge Discovery using our Darwin technology we accept as discriminators significant concept neighborhood differences.

Something of Imagery: The Web as a man-machine Teaching-Learning system



This figure shows us a vision of the Cyber space as a Teaching-Learning coupled system. We have two actors: users and authors that may interchange roles. The Web is represented by the circle where documents are hosted inside. There exist a restricted and always evolving EK, Established Knowledge. Users may navigate by the Web space thru Search Engines that may be functionally imagined as the exterior light green circle behaving like a **World Search Membrane**.

Users may query the whole Web obtaining information plus some subtle forms of intelligence. They may query either conventionally or semantically. This second form is not yet enabled but it will be shortly. Conventionally queries follow a sort of random paths as it is depicted in white inside the circle. Semantically they could point directly to EK and optionally to the rest of Not Yet EK pages, the Web majority. EK may evolve continuously. The search membrane could be enabled to register all world queries, a crucial data asset because it keeps hidden inside but retrievable “**Users Behavior Patterns**”! And at large enabling the knowledge of the **People’s Thesaurus**, in fact how people learn!.

On the “other side” authors injects more information and knowledge to the Web in a World Teaching role, and receiving as counteractions “traffic information”, querying patterns and direct users interactions under the form of demands, suggestions, registrations and even offers. We may also imagine a virtual **EK Membrane** that may control in the near future the best teaching and the best EK evolution.