Darwin
Distributed Agents to Retrieve the Web Intelligence
-For Web type Reservoirs-
Direct Search Engines: YGWYN-IOOC ([1])

Dr. Juan Chamero, Madrid, Spain 27th of May 2008; Reviewed at Buenos Aires, Argentina 22th of February 2009
Informal Draft addressed to Knowledge Management professionals

Something of History

At the end of year 1999, monitored from US we performed a sort of Database Search championship with students in their final of the ***Systems Engineering*** career in the ***Instituto Tecnologico de Monterrey***, from ***Mexico***. The championship consisted in querying a huge experimental industrial database, holding about 10 million firms profiles created to serve Small Businesses of the Spanish speaking countries logistical needs covering production, produces, manufacturing, markets, standards, regulations, and export and import information.

The first experiment

To facilitate the interaction a bilingual English-Spanish facilitator interface was also specially built. The users in this experiment could be considered bilingual or at least endowed with a rather high mastering of the necessary technical English. The descriptive information of products was carefully checked by a team of experts in languages and industry. Astonishingly the "semantic match" was inferior to 1%!.

Several interpretations within a wide spectrum were discussed in that opportunity, since students behaving like too clumsy or using the experiment to mock us for fun till the hypothesis that queries were somehow wrongly expressed and/structured even if written in a correct Spanish as it was ex-post checked. We disregard extremes and monitored in detail as observers how students built their queries and matching each query with its corresponding search engine outcome we arrive to the conclusion that most queries were incorrectly expressed to "**semantically match**" the database cognitive units: **words, keywords**, and **concepts.** They used correct acceptations either in English or Spanish to define specific meanings, for example "tires" that in the database jargon pointed to firms that build and commercialize tires meanwhile within the students´ minds the primal idea of it associated to images they have (car tires, bus tires, etc.) was associated to words such as "cubiertas", "pneumáticos", "gomas", "hules", "rodado" and many others Spanish acceptations. As most meanings hosted in the database were in fact "**multiple words keywords**" the matchmaking between meanings universes in two languages (ultimately between two visions of the same reality) probed to be hard to implement because the huge size and ambiguity of the possible "**keywords synonyms' space**". Another fact was that the same keyword considered isolated, out of its context, may belong with different meanings to many knowledge domains.

At that time I became myself a search engine expert and a studious of huge text corpus structures like the ones hosted in the "**Web Ocean**" deepening about meanings and differences among words, keywords, meanings, ideograms, ideas, concepts, subjects, and finally between information and knowledge. *"Keyword" was a "modal word" at that time that is a word or chain of words expressed in a given language and used by humans to locate, intercept and catch specific pieces of information (supposedly meanings) out of huge data reservoirs in order to satisfy their information and knowledge needs.*

Sometimes we found what we need in only one click but however most times if looking for something complex and/or particular we could spent hours in an unfruitful and many times misleading search. And even for expert users retrieving of some existent content probed to be almost impossible like if that content would be practically inexistent.

The second experiment

We performed with the same students group another complementary experiment. Students were challenged to retrieve topics (the resource at hand was a pool of conventional search engines) generated at random in both languages English and Spanish trying to do their best and not quitting the search until "up to them" something reasonable "satisfactory" was attained. All queries were

analyzed to find those potential "satisfactory" search patterns. What we found this time was a primal idea of "**modal keywords**". Let's imagine now using Google like a black box. Let´s think about a given meaning and start querying with a keyword at a time, preferable multiple words keywords, associated in our mind to the same meaning- between quotation marks. Take note of number of references, namely how many pages have within their text corpuses that specific keyword. We are by de facto immersed in a process of "modal keyword" discovering for a given meaning, grossly the one that brings more references (it should be the one that points better to "qualified" references). What we discover is that the semantic "tuning" of these modals is very sensitive, small variations (perhaps considered as negligible for humans) in their text form may lead us from too much references to none and vice versa*. In brief summary we acquainted that for a given cognitive subject, theme or topic, the modal key works like a magic key that opens for us the Pandora Box of a particular piece of knowledge "hidden" in the Web space*. Going back to the students' trials our conclusions at that time was that what they have found along their trials were approximations to modal keywords of the **Web as_it_is**, another important concept: semantic "baits" to retrieve specific pieces of knowledge extracted of the Human Knowledge as it is in the Web at a given moment. We invite you to see in our White Paper about "**Semantic Search**" the philosophical approach to these concepts.

Darwin prototypes

With this idea in mind at the time of the beginning of the Internet bubble collapse we decided to map at least a single discipline out of the Web as_it_is at that moment (year 2001). It meant a map enabling users to directly retrieve any piece of its related knowledge whether existent in only one click. By any piece we meant from a children looking for information about "Tiahuanaco" or "Machu Pichu" the lost city of Incas till a scholar looking for material to prepare his/her thesis about the Human Genome. To accomplish this task an Agreement was signed between ***Intag, Intelligent Agents Internet Corp*** an American R&D firm of which I was the principal and the ***CAECE University from Argentina. DARWIN Team***, a team of 20 stable programmers and content experts, was created to carry on this task. Intag conceded to this group the free use, for academic and related nonprofit purposes, of its proprietary **Darwin technology** that stands for **Distributed Agents to Retrieve the Web Intelligence**. This team was reinforced at testing time by the volunteer effort of more than 100 students and professors of CAECE University. As the final purpose of the Agreement was the creation of a **Map of the Human Knowledge**, something equivalent to the **Human Genome** in terms of Knowledge genesis and evolution, the Administration of the Agreement decided to begin with mapping something significant enough to draw conclusions about the final purpose feasibility: **Computing**, something well known and rigorously defined at that time, and perhaps equivalent to the historical task of mapping the **fruit fly genome** in the scientific journey towards the Human Genome. This first prototype was completed and announced in July year 2003 by the BBC of London in its Spanish Science Website and also presented in two International Congresses about advanced information technologies and in Universities of Argentina, Spain, Mexico, Ecuador, Chile and Peru. .

First Prototype description

Our first prototype was relatively easy to unveil because we knew how its "**logical skeleton**" would looks like. Computing was a relatively new discipline, very and well structured and whose "**topic names space"** had a strong consensus. We had at hand as a "semantic seed" to start mapping the "**Logical Tree**" published by the ***ACM Magazine*** with nearly 700 branches and a reasonable good glossary of terms, the ***FOLFOC Glossary*** with more than 12,000 meanings. Darwin agents unveiled from the Web a logical skeleton of Computing as_it_is in the Web of 1200 branches and a **Computing Thesaurus** of 54,000 meanings.

However we were still in the beginning of the beginning of knowledge unveiling via agents. In this prototype we had a semantic seed and a very structured discipline that fitted perfectly well to a logical tree, very specific meanings with negligible semantic fuzziness and noise. Our next step should be trying to unveil a more complex, ample, with great ponds of ambiguity, fuzziness and noise. In fact a discipline which enforcce us to start mapping without a feasible semantic seed,

namely from "**zero ground**" in terms of previous knowledge. Disciplines with these characteristics are abundant within the state of the art of knowledge: art, history, games, terrorism, sports, social sciences, etc. As an outcome of our first prototype experience we learnt how to unveil modal logical skeletons (logical trees), modal keywords, authorities, and how authoritative documents are usually written, something we defined later as WWD's, Well Written Documents formulae.

Second Prototype

To solve this puzzle take us four more years. In the beginning of year 2007 we were in conditions to intent mapping **World Art** staring from zero ground, meaning total ignorance, maximum level of uncertainty. In numbers this map displays over a modal logical tree of 7,570 nodes (we acquainted that nodes are semantically more meaningful than branches) distributed with a strong asymmetry along 13 hierarchic levels and pointing to an **Art Thesaurus** of nearly 300,000 meanings or concepts. A concept could be for instance a "work" like the Rigoletto Italian Opera of Verdi, a painting like the Monna Lisa of Leonardo, a subject like "street art", a work like a famous "graffiti" created by Japanese students, or a Ribera mural.

Uses of Knowledge Maps

You may ask now what this for?. And even more, many people believe that these maps exist somewhere in the Web!. This illusion is due to the incredible and varied offer of IT&C market. Most people get fun gaming and navigating thru open and free amenities and talking specifically of searching most people do not care if something they were looking for is not found. Let's suppose that someone is looking for the meanings of "The son of God", believe me a not trivial matter. If he/she does not find good enough references surely his/her attention will be probably captioned by other alternative meanings like for instance a musical play, or a collective farm, or the name of a workshop in Italy. And let's suppose now that we want to know all existent Art Maps. Thinking logically and as a function of our personal knowledge we may trust in well known "by de facto" authorities like for instance: Le Louvre Museum of Paris, El Museo del Prado of Madrid, The Metropolitan Art Museum of New York, the Guggenheim Foundation and many others. Of course all have their own visions of art and they are world authorities as well. *But Art has more than one billion of documents dispersed in the Web Ocean and perhaps an ideal "weighted modal" of art as_it_is expressed today in the Web is far from those visions!.*

By first time we humans have an open and free universal reservoir of all documented ideas and opinions, and we could dare to say that for each imaginable subject we may retrieve from hundreds to thousands of documents, and that almost nothing that deserve to be known is missing. Maps enable us to journey thru a real world that was up to now almost hidden to our sight.

Another wrong belief is that actual search engines know keywords: even the most powerful search engines only index documents by words. Complementary some search engines offer some semantic but limited guides to users. This is a valuable collective service generally provided by volunteers that intent to classify what they consider authoritative. Important and necessary efforts but not enough to see the whole Web as semantically ordered. Some years ago some languages like HTML enabled Website content owners to define the main subject deal with in their pages and their keywords but very rapidly this facility has to be ignored by the search engines agents because the confluence of ignorance, lack of specificity, abuse and perversity. Search engines like Google adopted the simplest and necessary criterion: to index only by words. As we analyze in our White Papers this criterion is necessary in order to have at least a trustable and everlasting text corpus of each document, the raw data in order to probabilistically unveil via agents its meanings.

Let's play a little with numbers. We have today about 12,000,000,000 retrievable documents of which only 8,000,000,000 are indexed. Let's also suppose that all are expressed in a single language. Is not known how many different English "words" either well or bad written have classified as existent a search engine like Google but a probable number would be around 20,000,000. So it is conceivable that once documents are stripped off from images, commands and all type of editing and metadata what remain are 8,000,000,000 text corpuses, in fact equal amount of semantic creatures composed only by words as elementary semantic particles.

A summary of this Web Ocean could be depicted as a huge two dimensional array of 8,000,000,000 columns by 20,000,000 rows. These virtual arrays could be “queried” today in a few milliseconds from any computer connected to Internet.

What is then missing?

Meanings are missing!. Meanings are expressed by words but are not detectable by agents unless we teach them how. It’s supposed that for a language like English we may define about 30,000,000 of such meanings. If actual search engines would unveil these meanings an expanded array of 8,000,000,000 documents (as of today, this number grows at a pace of more than 10 percent yearly) by 50,000,000 words and meanings would offer to their users a sort of “platinum service” of the type **YGWYN IOOC**, You Get What You Need In Only One Click. What they would need is a “**Semantic Librarian**” to guide users semantically thru a built in Human Knowledge Map. This map will have all possible subjects existing within the Human Knowledge namely meaningful semantic sequences similar to genetic sequences.

A search example

The above mentioned Semantic Librarian will guide users thru dialogs of the following type:

USER: I want to know about SUCH THING;
DARWIN Semantic Librarian: Perfect, but we suspect that your SUCH THING was incorrectly written in English. Perhaps you mean one of these acceptations: a list follows…..;
You may now proceed to: a) query again; b) to insist in conventional search with SUCH THING as it was written; c) to ask for our guessing:
Questioning follows until user need is recognized and located under a specific knowledge domain as it would happen in a Library guided by an expert librarian. This “location” implies the knowledge of the precise semantic chain where the man-machine agreed meaning belongs.
USER FIRST CLICK: If SUCH THING was for example Rigoletto the outcome of the guiding after a few man-machine interactions before issuing the first click would be something like:

Then you mean The Italian Opera Rigoletto within the semantic chain: Art => The Art => Performing Arts => Main Performing Arts => Theater => genres => Opera => Opera History => Italian Opera => Bel Canto movement => Rigoletto

And once the first click is activated the search engine will deliver the best authoritative pages dealing with that specific subject. From our research -up to now limited to our two prototypes- the probability of getting something valuable whether existent is around 99%. Darwin enables more clicks focusing in Markovian semantic similarities around first click subject neighborhood: up, down and collateral nodes.

Some References

Web

- Dr. Juan Chamero profile here: 1. http://www.intag.org/downloads/;
- Logical Art Map sample (Theater) skeleton, exported in Excel. Make click here: 2. http://www.intag.org/downloads/Z_theater.xls (¡It has three sections!);
- Collection of 20 White Papers about Darwin Ontology and Algorithms: 3. www.darwin-ontology.org;
- A Collection of 50 Darwin White Papers, You have to register first. It's free
- Darwin Home, the Darwin Team institutional Website.
- E-membranes to detect Users' Behavior Patterns, 4th WSEAS Int. Conf. on INFORMATION SCIENCE, COMMUNICATIONS AND APPLICATIONS (ISA 2004) Miami, Florida, April 21-23, 2004. You may download a free copy of it here: http://www.intag.org/pages/WP/
- Towards a New Digitalized Knowledge Paradigm. Presented at WACRA, The World Association for Case Method Research & Applications, 21st International Conference - Buenos Aires, Argentina, July 4-7, 2004. You may download a free copy of it here: http://www.intag.org/pages/WP/
- How Case Studies Methodology embeds with continuity within the millennial Teaching Learning Paradigm, Some reflections In opportunity of the Plenary Session of WACRA Congress at Buenos Aires – Argentina, July 6th 2004 and motivated by its main subject: "Cases as a Component of a Person's Research", presented by: Dr. Ronald Patten, De Paul University, Chicago Illinois, USA; Dr. James Camerius, Northern Michigan University, Marquette, Michigan, USA; Dr. James Erskine, the University of Western Ontario, London, Ontario, Canada; Dr. William Naumes, University of New Hampshire, Durham, New Hampshire, USA. You may download a free copy of it here: http://www.intag.org/pages/WP/

Scientific backup

Apart from the CAECE University from Argentina, Darwin Project has received the scientific backup of a Linguistic-Mathematic team coordinated by Dr. Eduardo Ortiz, Emeritus Professor of Mathematics and History of Mathematics of the Imperial College of London. They agreed to review the Darwin Ontology validity, and the semantic quality and reach of the first Mathematics Map extracted of the Web as_it_is by Darwin Technology.

Note 1: **YGWYN-IOOC**, stands for **You Get What You Need**: **In Only One Click**.