Darwin Process

As of November 17th 2007

Introduction

Darwin process is now an industrial process, a sequence of serial and parallel processes performed by agents but still managed by humans. Darwin is far from being a unique magic algorithm: the problems Darwin tackles are too complex, fuzzy and noisy for that. We prefer to talk of an “anthropic” algorithm that tends to become an autonomous multi agent industrial process instead. If we tag agents tasks by A and human tasks by H in the actual Darwin Block Diagram of about 80 boxes you will find today 70 A´s and 10 H´s. H tasks are performed via scripts, utilities and macros, never manually. I will try to explain you the logic behind this strategy.

Let’s go back to main frames programming of classical applications, like for instance Billing or Bill of Materials Explosion. To make a billing program for a small Corporation that sells a few lines of finished goods used to be simple, demanding no more than three to four months of programming. However if we are talking of a Billing System for a Corporation like IBM, Ford, Siemens or Toyota things are completely different, basically because problems derived of scale, variety and volumes to handle. These systems are to be designed having in mind Systems Theory and The Theory of Complexity. I would add a third factor: Errors. In small systems errors are few meanwhile in large scale systems are counted by thousands and derived from errors we have the problem of error propagation and error correction. In summary, large systems are at large limited by our ability to manage errors.

Let’s focus our attention to Semantic Trees: one thing is to manage trees of 100 nodes distributed along four to five levels and Web Semantic Trees of 300,000 nodes distributed in up to thirteen levels!. Human brains are not good to manage trees. Humans are rather clumsy and slow to locate errors in trees besides. Take into account that an error in a single node code could make the whole tree where it belongs useless and believe me humans are bad to locate these types of errors. We humans are good for Gestalt tasks whether trees could be envisioned within our visual field but large trees are split in thousand of tables and fields. Large trees should be created and loaded without a single error because they will have to be used by agents with zero tolerance to errors. And to make things worse Darwin raw data is extracted from the Web the largest, fuzziest and noisiest existent data reservoir. Notwithstanding Maps synthesized from the Web must be 100% precise, syntactic and semantic error free!.

Darwin technology is based in a set of 10 Conjectures but there is an upper conjecture above it, a Common Sense Conjecture not easy to test. We stated it as a hypothesis. Most research and findings are based on hypothesis that are forgotten if findings derived from it are successful and considered true. For example The Relativity Theory is based on the concept of simultaneity that only exists locally, as defined by local clocks. As the theory succeeded and how!, many scientists forget that basic assumption. What would happen if we can not probe that simultaneity exist at least locally?. What would happen if someone demonstrate the possibility of that eventuality?.

Darwin Basic Conjecture

People used to write coherently. This conjecture is in its turn based on “people” as a large enough statistical sample of “authorities” that thru “authoritative” document writing represents and determines the “Established Knowledge”, something that we give by certain in the Web Ocean. We assume then that this “Established Knowledge” is represented by Trees. So our basic and strong conjecture is that the inherent intelligence of any human discipline retrieved from the Web is or tend to adjust as much as possible to a math tree. Our assumption is based in the evolutionary sequence chaos, ideas, fuzzy logic, logic, and math. Math is perfection and all efforts made to approach to it deserve applause. The economy at small scale is chaotic, particular and sometimes represented by intricate graphs, meshes. The economic interrelations at country level approaches to a tree. Finally macro economy, at large regions and World level may be meaningfully represented by trees.

One of our 10 conjectures is that authoritative documents tend to be written managing concepts and keywords following certain logic rules we define as WWD, Well Written Documents. In our two prototypes we have tested that these rules apply perfectly. We suppose that a sort of evolutionary process controls the form human knowledge is established resembling the “invisible hand” that guides the economy, pumping up -in popularity- well written documents, inducing that keywords with high consensus tend to be used more and more because there are issued by authorities and at the same time become modal initiating a positive feedback cycle.

A Global View of Darwin Methodology

Another reflection is that searching for the best trees “hidden” (not easy to unveil) involves objectivity. Authorities (that should be discovered/confirmed) tell the world their truth, however a subjective truth, one among many. We attend all them and our agents weight their influences accordingly. So we may imagine the following feasible methodology (very globally) to retrieve the hidden intelligence for a given discipline.

Actors

H: Humans
A: Agents
W: Wizard scripts
P: Special scripts: programs and algorithms

Interaction Nature

C: Content Expert intervention
S: System Expert, System Architect intervention
L: Logic process
M: Math process
L-M: logical math process

a) MAPPING

1. H (C) Humans provide a seed of Authorities
2. A (L) Agents generate Virtual Communities of Authorities
3. H (C) Humans confirm Authorities
4. H (C) Humans provide a seed of Semantic Index
5. A (L) Agents unveil Authorities: retrieve pieces of intelligence and/or sub trees;
6. A (L-M) Agents edit these pieces of intelligence; order and correlate sub trees;
7. H (M) Humans define parameters for a meaningful Combinatorial Analysis to build Possible Trees
8. A (M) An Agent driven Optimization algorithm computes the most coherent combinations:
9. H (S) Humans approval;
10. A (M) Agents starts a preliminary logical math analysis of the coherence of the selected tree (see 12.2);
11. H (S) Humans accept or reject (rejection involves a second run);
12. A (L) Agents proceed to tree harmonization:
    - 12.1. A (L) Names consistency
    - 12.2. A (L-M) Attainment of the Specificity Rule: up-down-collateral
    - 12.3. A (M) Path Coding unification and standardization
    - 12.4. A (M) Ancestry uniqueness test
    - 12.5. A (M) Suspected virtual nodes detection
    - 12.6. A (L-M) Suspected redundancies detection
    - 12.7. P (L) Sort tests
    - 12.8. P (L) Tree by level
13. A (L) Agents proceed to Raw Data Map Search Engines Loading;
14. H (S) Humans confirms redundancies and perform minor adjustments;

b) WIZARD

1. A (L) Agents split Trees in Wizard levels;
2. W (L) Wizard script;
3. W (L) “Alice and Bob” AI script;
4. A (L) Alice and Bob alpha and beta tests;
5. P (M) Trade off analysis;
6. A (L) Agents build complementary resources: Authorities, Glossaries, Thesaurus access, etc;
7. P (L) Statistical script to guide Map updating;
8. P (L) Updating script;
9. A (L) Agents test going update;

c) WEB THESAURUS

1. A (L) Download Top references;
2. A (L-M) Referenced Pages Stripping;
3. A (L) Images Collection:
4. A (L-M) Text Parsing;
5. A (L) Word sample compilation;
6. A (M) Combinatorial Analysis;
7. A L-M) Common Words & Expressions filtering;
8. A (M) Potential Keywords classification and weighting;
9. A (L-M) Sample Doubling Iteration;
10. A L-M) Subject Core Keywords;
11. A (L-M) Core Keyword specificity test;
12. A (L) Thesaurus compiling;
13. P(L) Thesaurus editing;
14. P(L) Thesaurus Complementing Wizard;

d) DOCUMENTS PROFILING

1. A (L-M) Document Stripping;
2. A (L-M) Document Parsing;
3. A (L-M) Keywords discrimination;
4. A (M) Thesaurus Matching;
5. A (L-M) Document abstracting;
6. A (L-M) Document Profiling;

e) MAIN USERS BEHAVOIR PATTERN INFERENCE

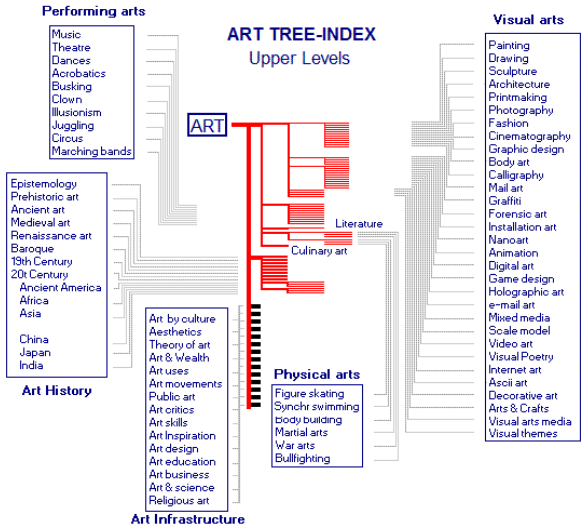
1. A (L) Users queries collection by sessions
2. P (L-M) n-ads combinations
3. P (M) patterns classification of queries
4. P(l-M) patterns classification by insistence and frequency
5. P (L) potential main behavior patterns
6. H (S) Human inspection
7. A (L-M) intrinsic and semantic popularities of chains
8. P (L-M) Potential n-ads People Thesauruses
9. P (L) Potential PAG´s, People Affinity Groups
10. A (L) Agents report of PAG´s activity

Steps Meaning

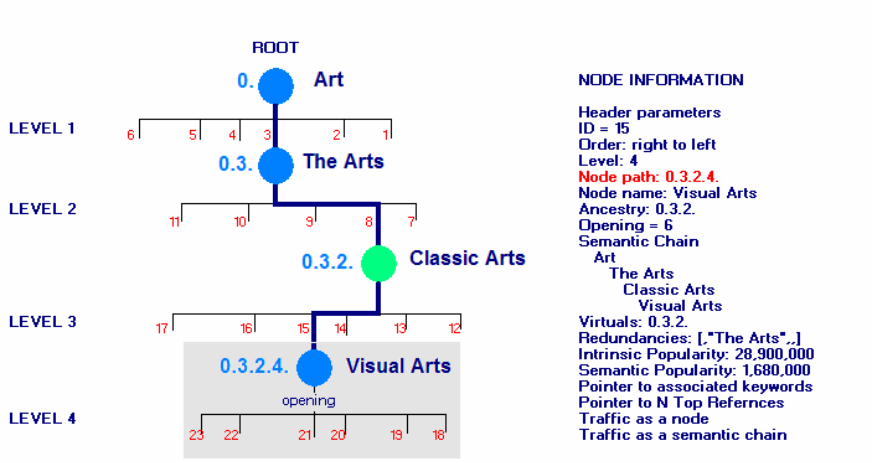
a) **Mapping**: A major Discipline (or all of them) of the “Established Human Knowledge” has being mapped **AS_IT_IS**;
b) **Wizard:** An “Alice and Bob” AI facilitator enables a **Semantic Super Search Engine at Thematic Level** build up;
c) A full **Web Thesaurus** enables a **Semantic Super Search Engine at Keyword Level**;
d) **Documents Profiling**: The next step. Documents being abstracted (Fingerprinted) when registering;
e) **Main Users Behavior Pattern Inference**: Main Users Behavior Patterns could be inferred from massive man-machine interaction **AS_THEY_ARE** without interfering with them!.

How is the inner structure of a Map?

The Art Map has 7570 nodes. Seen from above –upper three levels- it will look like the figure



Its skeleton is a logical tree. A small piece of it is depicted below:



Opening from root to leaves in up to 13 levels. As a database table it has about 20 fields, with a row for each node. In the figure order is ruled by ID as the Primary Key. The filed of ordering is Node path (in red) for a tree ordered “at path mode”. If field of ordering were Level we would have the tree ordered “at level mode”. The chosen ID order is from root to leaves and from right to left. Following this “arbitrary” ID ordering the field “Node path” is determined. Node names are determined by a specific script (the most popular synonym retrieved for each theme).

With this basic skeleton agents determine the ancestry and the opening of each node. Then agents proceed to determine the semantic chains from root for all nodes. These semantic chains must be processed before becoming queries. In the example is shown the chain “art” + “The Arts” + “Classic Arts” + “Virtual Arts” where node 0.3.2. (“Classic Arts”) is virtual and as such ignored for the query. As we will see soon virtual nodes are defined to harmonize the tree but their existences are not yet established. The chains need now to be cleaned from redundancies Within “art”, “The Arts” is a valid concept it is not a virtual node but real. However its presence will spoil the query efficiency because redundancy. Finally the query becomes: “art” + “Virtual arts” and the popularity for it will define the field “Semantic Popularity”. Redundancies are performed by a special script and approved by humans.

Intrinsic popularity is the one that corresponds to each “node name”. Results retrieved from queries are saved in another database. Each node will have pointers to these results and to their corresponding set of keywords. When maps are activated for running as cores of a semantic search engine some other data pointers are saved such as traffics: node traffic and semantic chain traffic for each node. Finally each node save a Header with the following information: pool of search engines used, date, number of Top References required, number of Top References retrieved, language, agent interface used etc.