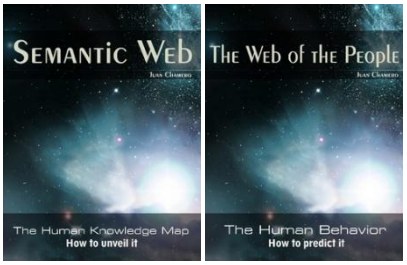
DARWIN “semantiza”la WEB

SU EVOLUCIÓN ONTOLÓGICA 2001-2017

Juan Chamero, Arquitecto Principal de la Metodología Darwin



Darwin es un proyecto de Inteligencia Artificial que lleva ya casi dos décadas tratando de interpretar y ver la Web en su verdadera magnitud, como semántica y siempre acelerada, corriendo tras nuestra propia sombra. Finalmente logramos crear una ontología apta tanto para agentes y robots como para humanos.

La Web Hoy

La Web hoy es una inmensa nube del ciberespacio que hospeda a más de 40.000 millones de Páginas Web, no estructurada semánticamente, que crece caóticamente a un ritmo del orden del 20% anual y cuyo contenido solo está indexado por palabras bajo precario control público.

Al presente acceden a ella unos 4.000 millones de usuarios, uno de cada dos habitantes de nuestro planeta, para comunicarse entre sí, dejar testimonio de sus vivencias y para buscar información y conocimientos esenciales para su evolución y supervivencia.

Situación actual: La Web no aparece estructurada semánticamente; sus hoy 40.000 millones de páginas Web no están preparadas para ser leídas por agentes a la búsqueda de una “temática” dada. Hoy solo es posible recorrerla a través de buscadores “saltando” de una palabra a otra o combinando palabras remedando conceptos, inspeccionando y leyendo pedacitos de contenido sospechado de estar relacionado con la temática buscada, a solo juicio del humano que la explora.

Se Propone

Ver más y mejor la Web cual si su TODO estuviera perfectamente estructurado semánticamente.

Para por ejemplo conseguir en un solo clic lo que buscamos o pistas seguras para acceder a TODO lo existente en la Web sobre el tema en cuestión (a la larga… sobre CUALQUIER COSA).

¿Cómo? Detectando, accediendo, recuperando, **“v**iendo” y asimilando temáticamente el contenido Web tal como se muestra a un momento dado.

Cabe preguntarse: ¿Pero esto no existe ya? ¿No está el contenido Web clasificado temáticamente? ¿No tendría la clasificación temática que haber sido hecha desde sus comienzos por los propietarios de su contenido? **¡La respuesta es No!**, cada “Página Web” está subida tal como fue creada sin “metadata” alguno, como por ejemplo los libros en una biblioteca pública, que posee para cada uno una tarjeta con suficiente detalle semántico como para evaluar su confiabilidad y potencial cognitivo.

La **Web_tal_como_es**: Así, encontraremos en la Web contenidos estructurados de perfecto a pésimo o nada, de bien a mal escritos con y sin errores ajustados a verdad o a engaño y ya alcanzando volúmenes enormes operando en escenarios del tipo Big Data.

La Web como legado de la Humanidad: Por otra parte el contenido Web es y debe seguir siendo abierto, accesible e inviolable es decir cada uno de nosotros debe poder acceder a cualquier contenido pero solo para "leerlo" y/o "bajar" lo permitido. Luego en teoría podría plantearse la opción de leer totalmente su contenido, del orden de 40.000 millones de páginas Web, para enterarnos de su temática y eventualmente clasificarlas en una base de datos construida al efecto. Esto aparece a priori como una tarea crucial y necesaria aunque de difícil y costosa realización. .

Solución al dilema de una Web supuestamente semántica que no lo es: El **Darwin Team**, conformado por expertos en **Inteligencia Artificial** y en **Sistemas de Alta Complejidad**, creó la **Metodología Darwin** basada en una Ontología orientada a emular probabilísticamente como documentamos los humanos. Se comenzó a aplicar dicha metodología a universos muchos más pequeños y limitados aunque igualmente enormes, del orden el 1% de la Web,encontrándose que curiosamente los humanos documentamos en formas bastante estructuradas, más de lo imaginado aunque diversas, informales y desde un punto de vista probabilístico.

La Web como reservorio de Explicaciones: La idea clave que abre las puertas a la recuperación no solo de datos e información sino de conocimiento y hasta esbozos de inteligencia es lo que la Metodología Darwin denomina "**Especificidad Semántica**":

……los humanos tienden a explicar y consiguientemente a documentar sus "ideas mentales" ("in mind ideas") creando y/o empleando literariamente conceptos y neologismos específicos de la temática que intentan explicar!

La explicación logra su cometido generando contenido comprensible combinando en forma inteligente palabras y expresiones comunes de una "jerga" dada con conceptos del mencionado **Conjunto de Conceptos Específicos**. Para ello Darwin extrae de la Web manojos completos de referencias, brindadas por buscadores convencionales como Google para temáticas dadas, digamos de 10.000 a 40.000 documentos cada uno y de cada uno de ellos detecta y recupera, en función de su "rareza", su respectivo Conjunto de Conceptos Específicos.

¿Qué se necesita además para ver más y mejor a la Web?: Si somos capaces de detectar y recuperar de la Web a sus criaturas cognitivas, coloquialmente "pedacitos" de conocimiento bajo la forma de conceptos, deberíamos ser capaces ahora de detectar y recuperar absolutamente a TODOS y cada uno de esos pedacitos en forma estructurada, es decir sujetos a disposición y orden dentro de un TODO.

El **Conocimiento Humano** en la Web: Al respecto los humanos hemos expresado a lo largo de los siglos una marcada predilección por la estructuración arbórea, fundamentalmente bajo la forma de "**Árboles Lógicos**" invertidos quizás queriendo recuperar parte de nuestra libertad y pertenencia cósmica perdida. La Ontología Darwin supone que a estas estructuras no hay que recrearlas ni inventarlas ni siquiera sintetizarlas sino que subyacen escondidas en el espacio Web bajo distintas formas: índices, tablas, grafos, glosarios y ¿por qué no? hasta como "Tesauros".

Darwin a lo largo de un proceso de "**destilación de conocimiento**": detecta y recupera estas estructuras arbóreas mediante un complejo proceso recursivo de más de 90 pasos partiendo de "**Conocimiento Cero**" y/o haciendo crecer "**Semillas Semánticas**" algo así como índices abreviados supuestos para cada "**Rama del Conocimiento Humano**.

TESTS: Complementariamente, Darwin posee algoritmos que evalúan para cada rama y para cada estadio real o virtualmente evolutivo de la misma la coherencia lógica y probabilística de la hipótesis arbórea. Finalmente estas estructuras y criaturas pueden actualizarse total o parcialmente en forma autónoma y de hecho perfeccionarse ajustándose más y más a la realidad.

Conclusiones preliminares: Mediante la aplicación de ontologías semánticas como Darwin orientadas a emular como pensamos y documentamos los humanos se hace viable detectar y recuperar de la Web datos, información, conocimiento e inteligencia suficiente como para crear en forma cuasi directa **Idel's**, **Informes de Inteligencia** y **Mapas del Conocimiento** de alta precisión y confiables sobre prácticamente CUALQUIER COSA (o TEMA) del TODO humano.

Referencias y Epílogo

¿La indexación por palabras es necesaria?¡Sí, es absolutamente condición necesaria mas NO suficiente! Google cumple bien esta especie de tarea pública global indexando no solo todas las palabras documentadas, para todo par lengua cultura, sino además enormes encadenamientos de palabras tales como gruñidos del tipo grrrrrrrrrr y millones de frases célebres, poesías, coplas, etc.

¿Qué es lo que faltaría entonces? Faltaría **"mapear"** la totalidad del conocimiento disperso en la Web a un momento dado: tarea de alta complejidad pero hoy viable y accesible.

Bibliografía específica Darwin

Mapa del Arte (2010): solicite una presentación a Juan Chamero: jach_spain@yahoo.es

- CV Académico y Profesional: http://juanchamero.com.ar/jach_cv_professional.pdf
- CV Humanístico: http://juanchamero.com.ar/jach_cv_personal.pdf
- Mapa del Conocimiento Humano, primer prototipo, BBC de Londres (2003)
- Semantic Web,Lulu ebook (2012)
- The Web of the,People, Lulu ebook (2013)
- The Web: A World of Avatars, Google ebook (2016)

Estructuras arbóreas y algunos enlaces sugeridos por Darwin – de Platón al presente en orden inverso

- Conocimiento estructurado bajo forma de árbol;
- Grafos de Conocimiento según Google;
- ¿Qué hace Google? Es una especie de Notario Público Global obligado de certificación del contenido Web, su existencia y validez así como de su importancia y popularidad.
- Estadísticas Web
- Cantidad de palabras en todas las lenguas
- ¿Cuántas palabras debo conocer? Universo de palabras – Regla 95/5
- Semantic Web Standards as per W3C Consortium
- Semántica como estudio filosófico del significado
- Acrónimos, abreviaciones, neologismos y barbarismos en la Web
- Conceptos y Cuantificación de Conceptos
- Platón y su Mundo de Ideas

Darwin está documentado en ~2.000 páginas PDF y Word escritas en inglés y en español. Se recomienda explorar previamente contenidos actuales científico filosóficos sobre: idea y concepto, datos, información, conocimiento y sabiduría, inteligencia y mente.