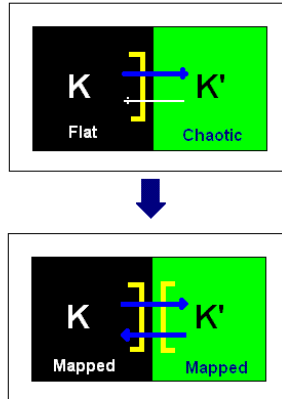


How to Manage the Human Knowledge

A set of Conjectures

Juan Chamero, CEO [Intag](#), December 3rd 2003

Note: updated as of 20th November 2004 in fuchsia



Actual versus new Knowledge Management paradigm is depicted in the figure above. Established Human Knowledge K hosted in the Web space (black region) interacts with million of pieces of People Human Knowledge K' connected to the Cyber space (green region). Actual interaction proceeds broadcasting from K to K'. K' appears as chaotic meanwhile K appears as structured even though "flat" in only one level. People are only enabled to query K. Data and Intelligence only flows from K to K'. Under a new paradigm that uses Internet at its full capacity a win-win matchmaking scenario could be expected optimizing the overall Human Knowledge. Throughout an e-membrane (yellow) Expert Systems aided by Agents may map K Realm first. Once K is mapped K' becomes a matter of understanding as seen from K side and Agents may proceed to map it. Once both sides are structured and well known, intelligence, understanding and even wisdom may flow from K to K' and conversely from K' to K.

Abstract

We have created a methodology to manage the HK, Human Knowledge. We distinguish two types of HK, the "Established" HK, normally structured, and the People's HK that looks like unstructured, chaotic, and dispersed. The methodology deals with these two realms trying to unveil, detect, and classify its main features. The core of the methodology is an Expert System aided by Agents that work with these two realms throughout an e-membrane that as a living membrane facilitates the open and free communication between them trying to create win-win scenarios without perturbing the actors. The Expert System works based on a set of common sense "Conjectures" that are now in process of being scientifically tested by the AI-Lab of the [CAECE University](#) from Argentina. We state that it is possible to "retrieve" content and intelligence automatically from data reservoirs. As an example, given the Web we may automatically generate its "hidden" Thesaurus solving in this way the basic human searching problem. As long as people communicate openly and freely through an e-membrane their suspected behavior patterns could also be unveiled, detected, and classified. The conjectures set deals with a basic Conjecture: the triad Logical Tree, Thesaurus, and Cognitive Objects unequivocally identifies any type of Knowledge.

Our finding could be a milestone that points to social communications: on one side juridical and physical persons that sale, teaches, communicates, and people that buy, attend, communicate, and that continuously issue opinions about everything conceivable, on the other side.

We have created a first prototype that hosts a HK map on the "establishment" side and that has an e-membrane with full capability to enable the Expert System that administer the prototype to learn as much as possible from users and at the same time detects, and classify every suspected users' behavior pattern. The HK map evolves by itself. The Thesaurus is automatically generated from the Web by agents (in fact we got it for a prototype focused on ICT, Information, Computing and Telecommunications, see it at www.intag.org). The methodology could also be applied to retrieve the intelligence of huge historical data reservoirs.

Introduction

I will try to summarize the core concepts we were working, related to IR, KM, Web Intelligence, Web Matchmaking, Expert Systems and Agents. Everything started trying to optimize Web Matchmaking via Expert Systems aided by Agents. The more we advanced the more we realized that everything was closely related to the Human Knowledge understanding. The Web allows us the possibility to deal with a credible sample of it as a bulk and with its tiniest components, websites and pages at the same time. Our first work concerning the purpose of this document dealt with a huge American industrial and commercial database storing more than 10 million registers designed to be matched by millions of concurrent users. At that time and at that scale we realized of the existence of two realms, users and owners with an extremely poor matchmaking, more than expected based on their reasonable semantic differences. Our suspicion was: people “talk” and even “think” different when acting either as owners or as users. If that’s true perhaps it is a trivial fact. Not so trivial was to focus our attention to build a man-machine user interface between them capable to register any type of conceivable event at any side of the interface. We were in that direction because we were working with an Expert System, allegedly enabled to learn as much as possible from users. At that time we were astonished by the pronounced apparent asymmetry of the matchmaking itself: the “owners” were by “de facto” enabled to broadcast anything they wanted and now with an Expert System empowered by a sort of “e-membrane” they could be enabled to learn “as much as possible” from their users besides!.

We believed that in the long range this asymmetry will not be as good as symmetry for the overall health of the man-machine coupled system. Asymmetries determine win-lose scenarios which at large become bad for both, owners and users. We intuited that we were in the right track to arrive at a general solution to the general matchmaking problem.

Conjectures Set

Conjecture 0: the triad Logical Tree, Thesaurus, and Cognitive Objects unequivocally identifies any type of Knowledge

It is unchanged in its essence. However it implies a new Thesaurus definition encompassing the whole triad. In the beginning we accepted the notion of a “flat” Thesaurus as a single set of keywords for a given discipline. Then we upgrade the concept defining a structured Thesaurus with sets of keywords for each subject of a given discipline. Ultimately we may define a Knowledge Thesaurus that includes a basic library of meaningful documents for each subject with their corresponding keywords. This type of Thesaurus behaves like a “Knowledge Tree”. In figures let’s talk of Medicine. A typical curriculum would have about 800 subjects, 80,000 keywords, and 1,000 keywords per subject in the average. A typical Medical Library would have 8,000 basic documents. Of course this triad could identify Medicine, being the content core and the intelligence of let’s say a reservoir of 30 million documents dealing with Medicine.

Warning: in all these conjectures the crucial point to take care of is “keyword”, a representation of a meaningful concept in a given discipline, by symbols, acronyms, words and precise strings of words. For instance NIC, Bell, IBM, parallel processing, gray codes, cardiovascular disease, zen fitness, e-business, n-tier, are keyword examples. Humans used to document ideas combining and threading sequentially “common words” and keywords semantically following very specific syntax rules in documents dealing with specific “subjects”. Each subject is a keyword but the contrary is not true, most keywords are not subjects. Subject is a category of understanding and wisdom. In our knowledge paradigm we talk of the ascending hierarchy data, information, knowledge, understanding, wisdom. Subjects are crucial themes societies need to survive. The established order must provide them, in fact knowledge structured in subjects.

Most conventional Search Engines index by words instead of keywords. Some of them only offer, from time to time, lists of most frequent users' words/sequence of words, as highly probable keywords. They don't differentiate them neither by subject nor by discipline and they don't "learn of them" to improve the documents indexing. For this reason we say that they are "flat", unstructured.

Conjecture 1: Website Owners "speak" and "think" rationale in terms of their objectives and in terms of their matchmaking policies.

It means that they send programmed "messages" (broadcasted or at demand) to their users. Its thinking is usually represented by "Menus" of features, attractions, associated with a sort of "Logical Tree". They think as hunters preparing their hunting strategies. Even in the case of educational sites it happens pretty much the same: teachers have to succeed "broadcasting" and at large inoculating pre established knowledge in their students. The same happens concerning governors and their governed. From a simplified semantic point of view the owner's behavior is structured and thought as something fixed: the established True. This structure could be represented by a Logical Tree. These logical structures are well suited to describe the "offer", at large Cognitive Offer. If we were talking of Catalogs, for instance, the Logical Tree must describe them thoroughly, the main lines of products, their subdivision by type, material, functionality, until the leaves. When talking of Knowledge we deal with a logical tree of Major Subjects or Disciplines and going down to sub subjects until we arrive to very specific leaves.

We conjecture that the human being as a user, in the role of asking for something, is not well suited to ask for subjects belonging to K Realm. As we shall see later in the users' realm reign people's subjects instead.

This conjecture remains the same. We say rationale because they think along a tree and speak like automata answering users' queries like governors versus governed. As a matter of fact things in K Realm are "fossils": laws, regulations, codes, prescriptions, are the consensual truth at a given moment inherited from the past. Their destiny is to change to adapt better to changes in K'. In K things are frozen at a given time, for instance "as of ...". That's not bad!. It's the eternal game of evolution. In K life is continuously brewing, the new ideas, suggestions, opinions, all came from K.

Conjecture 2: Users "speak" and "think" rather chaotically in terms of their passions, desires, their necessities at large.

As a corollary, owners and users speak and even think different. A smart observer, looking from a virtual "e-membrane" at the owners' side, with analytical spirit, may arrive to the conclusion that any offer has a well defined purpose and that its "hunting" strategy could be precisely inferred. On the contrary, it would be practically impossible for any observer located in the same place to infer any type of order at the users' side without accumulating extensive man-machine interaction experience and being smart enough to use advanced computing techniques and bizarre Human Behavior theories. Users try to satisfy their curiosity, and ultimately their needs, gaming against the "other side" at their own risk with their talent, their knowledge, and with strategies changing from time to time and from situation to situation. Users speak with personal "keywords", a word or string of words –understood by K- that point to satisfy their needs. These keywords (or their components) must exist in the owners' side in order to succeed. For that reason the whole collection of possible keywords should be thought as created at the owners' side and organized in a Thesaurus, associated throughout a "non inheritance" rule to the Logical Tree. By non-inheritance rule we mean "keywords specificity", namely the right use of them at the right level of comprehension within a document. Let's imagine documents as been either too general or too focused. If too general authors should use general level keywords avoiding, as much as possible, the use of too focused keywords. The same could be the rule concerning too focused documents: authors should use focused keywords instead avoiding the use of too general keywords. Our algorithms have the capability to assign keywords to its level of specificity.

This conjecture remains the same. Users need "solutions", prescriptions that supposedly are stored and publicly offered in K Realm, in most cases openly and free. To get them the only tool they have at hand are symbols, buttons, links, and at large string of meaningful words of K side. So they are obliged to know precisely the K jargon to succeed.

Conjecture 3: Users' interactions along sessions are strings of semantic molecules of two types, users' keywords, and navigation instances. The sessions' strings are the representation of the users' strategies to satisfy their needs.

With these two conjectures in mind it is possible to fully satisfy basic users' needs as "predicted" at the owners K side. Well written keywords but not existent will cause a mismatch condition, and it's a warning for a potential evolutionary change to be taken at K side.

This split suggests two separate basic users' patterns: keywords' patterns, and navigation's patterns. Instances before and after a given keyword, slightly modify both the meaning and the result obtained via a given keyword. Concerning searching strategies the experience tells us that in huge clusters the keywords sequence - keywords may or may not belong to the same discipline- doesn't matter too much: [k1, k2, k3] and all its permutations could be considered similar strategies. A strong supposition is that the interleaving of instances within a session doesn't affect the strategies outcomes significantly.

For all these reasons users sessions' tracks of the form [iikikikiikii....], where k stands for keyword and i for instance, even ignoring their related outcomes, provide us a primal source to infer users' behaviors without interfering with them! (For instance, correlating outcomes to given strategies). The sequences of k n-ads, monads, dyads, triads, and so forth, detected in users' sessions, could be considered then like Tarzan's expressions of the users' Jargon.

Yes and not. All depends of what side you are in. A given word or string of words could be considered as a guess, the right "key" to open the Pandora box of K. If one user succeeds he/she may infer then that the key used is a keyword. From K side, you as an observer, may infer that a given word or sequence of words is a keyword if and only if it has been issued a significant number of trials by a significant number of users. Here we may envisage then two types of keywords, k and k' which stand for keywords of K Realm and keywords of K' Realm. As we will see we may also define two Thesauruses, the K Thesaurus and the "people" K' Thesaurus.

Conjecture 4: Cognitive Objects, documents, are expressed as strings of two semantic molecules, Common Words, belonging to a given Jargon, and keywords.

These cognitive objects are closely related to subjects of the Logical Tree, and this relation could be either explicitly or implicitly stated. We are especially interested in documents that behave as "Authorities" for each subject of the Logical Tree. An ideal authority should deal with only one subject of a given discipline and its keywords should be as specific as possible, that is it should use keywords that usually appear at the document subject level. It means from minimal to zero presence of upper and lower level keywords.

Note: Catalog' items, for instance "Buyers and Sellers Catalogs", should behave_as/correlate_to documents, are similar.

Conjectures 0 through 4 are closely interrelated. One thing astonished me when dealing in the past with newspaper marketing people was that they talk of the very "cognitive" content as "blank", only attractors because for them the newspaper could be viewed as a physical media to carry ads and businesses. So we may then imagine two visions: for regular users like you and me, the newspaper is a media that reduces our ignorance, our uncertainty, and ads, all kind of propaganda are blank, and only eventually cognitive objects, just the contrary that for owners!.

Conjecture 5: It is possible to enable a Full Duplex Type communication between Websites and their users throughout an e-membrane, enabling the free flow of content and its associated intelligence between them.

In our architecture all the interactions are performed through an e-membrane. This membrane resembles a living semi permeable membrane where all the messages going forth and back through it are processed trying to extract as much information as possible from them and at the same time providing the best service to both sides. Let' see what it means.

When the message is a potential keyword coming from users' side, the membrane performs all possible "static" statistics at both sides, namely: it accounts for Thesaurus keyword use at owner side, makes a request to offer all documents available, and accounts for the user history. Besides these statistics the keyword is analyzed and matched versus a set of Users' Thesauri. Each keyword within a session is potentially considered as part of a "speech" or part of a sort of interception game (a "Get the Truth" game), and accordingly it could be accounted in more than one Users' Thesaurus.

This e-membrane was thought as a smart, subtle, and polite mechanism. Its aim is to obtain as much information as possible from the users' side without interfering with them, or at least minimizing interferences. The core of its "smart" built in strategy is to keep it from spying users. No cookies, no brute "bail and catch" hunting devices. Total absence of owners' side messages such as: Come here again!, You are Welcome, and even the mild May I help you?. To offer an open and free interface the e-membrane is built to stop and/or discourage any type of intrusion such as: How was it?, Was it helpful to you?, and never issue of "offenses" like "Please tell us a little about you". The dominant policy is "let it be", "let it flow freely", concerning the communication between owners and users.

Warning: This conjecture remains pretty much the same. However we understand it now better. We may envisage hitherto what symmetry of K and K' realms imply. In K we understood the tree structure and pairs [k, s]. In K' we need to "understand" what the meaning of [k', s'] is.

Conjecture 6: Intrusions in communications cause serious troubles that go deeper and far than a local perturbation. The slightest intrusion may make invalid not only the session but prevent users from communicating freely. They distort the static statistics and the users' strategies as well.

This is trivial. More than a conjecture it could be considered a fact, something that has gained consensus. However these expected effects should be investigated and measured because there is a wrong belief nurtured by surveys and polls. Conventional surveys and polls are based on intrusions, sometimes generating visible harassment. These intrusions condition the answers. How much?. It has to be measured in order to eradicate it as much as possible as a credible methodology to know the K' Realm.

Conjecture 7—concerning Human Knowledge-: Human Knowledge is bounded.

We were talking about the practical use of HK in physical terms. One approach could be the volume of documents necessary to describe it. If we estimate the total amount of written documents in 12,000 millions –each one contributing with a piece of "valuable" information- we may ask ourselves then: How many of all those documents could be considered "Authorities"? We are also talking about the "Established" Human Knowledge; let's say the sum of all the structured knowledge as "Disciplines", the Major Subjects of the Established Human Knowledge. In this sense a Major Subject could be Medicine, Engineering, Philosophy, and the basic Curricula we learn in the universities (we have to add more subjects to the list, like for instance all derived from entertainment issues, the "homo ludens" activities).

Along the last decade there were some approaches to measure this Human Knowledge variety. The Britannica for example tells us about 8 super disciplines in the upper level: history, philosophy, art, society, technology, religion, science, and mathematics. Some of them open too much like technology, and science meanwhile some as religion, and mathematics open in a few branches.

We have found serious essays that talk of HK spectrum ranging from 150 to 200 Major Subjects. In its turn Major Subjects could be opened in Logical Trees from 600 to 1,000 subjects. Well let's talk now a little about Authorities.

We mean by authority a document dealing with approved/certified authoritativeness about a given subject. If the subject is up in the hierarchy the document could be physically organized as a Manual, or as a Treaty. Going downwards in the tree we may talk of Essays, and Papers. As you know this is not strict but a trend. Anyway we are talking of whole documents not single pages. This comment is addressed to the Web, where the unit of retrieval via search engines is the page. Authorities are generally hosted in Web sites not in single pages.

We estimate 3 authorities in the average per subject, in order to have a standard library for a given discipline. Talking in Web terms we mean 3 Websites per subject. You may imagine how a university professor ideal library would look like for comparison. For a discipline of 1,000 subjects that would mean 3,000 non repeated books to cover the whole spectrum. Not bad!. However this is not enough to be fully satisfied. We need more than that; we need redundant, complementary, and similar documents. Let's see how we could cope with this necessity.

The next step is to have the Human Knowledge Thesauri, the whole collection of the whole HK (from the "Established" side). Let's perform some calculations to estimate the whole size:

Major Subjects: 200
Subjects per Major Subject: 1,000
Authorities per Subject: 3
Basic HK: 600,000 documents
Extended HK: 200 million documents (see below)

We estimate that the average Thesaurus will have about 50,000 keywords, so we are talking of a Web Thesaurus: of nearly 10,000,000 keywords

Now let's go back to our need of more information to enrich our Basic HK. We have two options. First one is to retrieve similar authorities to the one we have at hand. Second strategy is to look for the best documents that satisfy our initial curiosity. In both cases our starting point is a basic document "a priori" considered good. How many documents could we retrieve with almost the same quality (authoritativeness)?. The answer is: as much as we decide depending of the existence of them and the quality resolution of the retrieval procedure. Our optimistic estimation is 20 per query, so extending the limit of our Extended HK to 200 millions of documents. Not too bad again!.

This assertion is especially worthwhile if we re talking of a current HK continuously maintained, updated with the news of it, and corrected. That means an evolutionary HK reservoir.

This is an ontological concern. Remember that our knowledge paradigm only make reference to the digitalized knowledge, formally expressed at a given moment in a given language.

Conjecture 8: Given a LT we may generate automatically its related TH

This conjecture was confirmed. Our first Thesaurus of 53,000 keywords for the ICT, Information, Computing, and Telecommunications discipline was initially determined half by humans and half by agents. With this experience we have designed and implemented a procedure to perform the automatic generation of TH given LT provided we have access to a reservoir where from to infer the semantic components of a given content.

Be that reservoir the Web and the LT of Medicine. We need a huge amount of suspected "Authorities", for instance more than 100,000, to apply our procedure. Once TH is generated we may proceed then to determine the Basic Medicine Knowledge and the Extended Medicine Knowledge.

Be that reservoir a proprietary database and the task is the engineering/reengineering of a Catalog. The same considerations apply.

This conjecture remains but it was poorly defined. It presupposes the existence of a huge reservoir of documents dealing with subjects belonging to LT curriculum. However this conjecture is hard to prove. We may envisage a procedure (in fact our Darwin-First methodology is one procedure) to perform that in a series of steps. From LT we extract a flat Thesaurus seed; with this seed we extract a first knowledge base; from this knowledge base we extract a whole flat Thesaurus (not a seed); with this flat Thesaurus we extract a better knowledge base; with this knowledge base guided by LT we structure the whole thesaurus, let's say we obtain all possible semantic pairs [k, s].

Conjecture 9: Given a Historical Reservoir we may generate its related TH and a collection of its main Subjects and Themes.

We are now pursuing this conjecture to become a reality. The crucial problem to solve is the automatic LT generation. We may close the cycle by human intervention as a first approach. Agents offer a set of weighted LT alternatives.

This is in certain way the corollary of Conjecture 8.

Bibliography

- 1 [ICONS, Intelligent Content Management System](#), Feb 2003, Rodan Systems (PL), The Polish Academy of Sciences (PL), Centro di Ingegneria Economica e Sociale (IT), InfoVide (PL), SchlumbergerSema (BE), University Paris 9 Dauphine (FR), University of Ulster (UK); It deals with Access Algorithms and Data Structures Underlying a Distributed Knowledge Base.
2. Web Epistemology, <http://webepistemology.org/main>, is a platform for studying the impact of the web on the mechanisms of collective knowledge production, organization and distribution. A major part of this project is devoted to analysing the stakes and goals of scientific research management and policies with regard to the web.
3. [Distributed Knowledge Modeling through the World Wide Web](#), by Mildred L. G. Shaw and Brian R. Gaines, from Knowledge Science Institute, University of Calgary, Alberta, Canada T2N 1N4, {mildred, [gaines](mailto:gaines@cpsc.ucalgary.ca)}@cpsc.ucalgary.ca,
4. Castells, M. (2001), "The Internet Galaxy – Reflections on the Internet, Business, and Society", London; Oxford University Press.
5. [Web Hunting: Design of a Simple Intelligent Web Search Agent](#), by [G. Michael Youngblood](#).
6. [KWIC](#), Keyword in Context, an old but always present Concordance concept and tool, as a parsing general utility.
7. Knowledge Management – [Digital Cloning \(KES\)](#), Digital Informational Experts Created from Prior Personal Knowledge, where KES stands for Knowledge Expert Systems. It's a reflection about knowledge, personal and collective, and what information and knowledge management are.
8. [Actualizing Context for Personal Knowledge Management](#), by Kenneth A. Berman and Fred S. Annexstein, Department of ECECS, University of Cincinnati, Cincinnati, OH 45221. ken.berman@uc.edu , fred.annexstein@uc.edu.

9. [KAON](#) is an open-source ontology management infrastructure targeted for business applications. It includes a comprehensive tool suite allowing easy ontology creation and management and provides a framework for building ontology-based applications. An important focus of KAON is scalable and efficient reasoning with ontology.

10. [Classification and Clustering](#), extracted from Kiduk Yang's Doctoral Student Thesis, [North Carolina University](#).

11. "[El Espacio Web y la Noosfera](#)", (The Web Space and the Noosphere), by [Dr. Juan Chamero](#) and published by SCIPPOOL–Red Cientifica, a bilingual Spanish-English scientific community, 2001.

12. "Towards a New Digitalized Human Knowledge Paradigm", published by SCIPPOOL –Red Cientifica, a bilingual Spanish-English scientific community. You may see it in [English](#) at and in [Spanish](#).

13. "Towards a New Knowledge Management Paradigm", ISSN 109-2750, presented at WSEAS, the World Scientific Engineering Academy and Society Multi Conference held in Miami, USA, April 2004, WSEAS Transactions On Computers, Issue 5, Volume 3, Page 1488. It's available for members at www.wseas.org and as an abstract in www.wseas.com.

14. "[How Case Studies Methodology embeds with continuity within the millennial Teaching Learning Paradigm](#)", published by SCIPPOOL –Red Cientifica, August 2004, a bilingual Spanish-English scientific community.

15. Presented as a New Generation Search Engines Tutorial in the Argentine Expocomm 2004 by the IEEE Buenos Aires Chapter, held in Buenos Aires, Argentina, September 2004, You may obtain a copy of it by asking for <http://expocomm.com/argentina/> or to the lecturer juan.chamero@intag.org.