The end of Conventional Search Engines

Juan Chamero, juan.chamero@intag.org, CEO Intelligent Agents Internet Corp, http://www.intag.org, August 31st 2004
Voice Contacts: US, T: 214 893 5010, Spain, Madrid 652078203 (mobile)

The American Research and Development firm, Intelligent Agents Internet Corp, has created an Artificial Intelligence Methodology to build "e-lenses" to see the hidden order in the Web that will enable Internet users to find what they need in only one query. To demonstrate its feasibility they have built a prototype of these lenses for a single discipline and measured the efficiency of searching if guided by encyclopedic semantic skeletons such as Britannica for British English and Encarta for American English that at large induces semantic styles. The experiences performed demonstrated that these e-lenses that could be automatically built and cloned by intelligent agents for each language and for each discipline enable users to find what they need in only one query. The prototype has been the result of a joint effort with the CAECE University from Argentina (http://www.caece.edu.ar). .

These e-lenses will induce a Web revolution not only in the search engines industry but in the Web uses as a whole. What does "one click to find what you need" mean in this context?. Million of users waste their time looking for information without finding it. In a near future throughout special e-lenses millions of users, common people, students, professionals, employees, will be able to create and host in their personal computers their own knowledge bases and the intelligence they need to the efficient fulfillment of their daily activities, by extracting them out from the Web instead.

Explanation

Conventional Search Engines are inefficient in retrieving knowledge and are near collapsing due to the exponential explosion of documents and by unscrupulous marketing procedures. Even the most advanced ones are unstructured, flat, where billions of pages are indexed by ambiguous pseudo "keywords". Consequently, answers to queries provide myriads of references belonging to dozens of different disciplines because each of these pseudo keywords may represent dozens of different concepts belonging to an equal number of disciplines. The ambiguity of single word keywords like "flesh", "core", and "thread" when their meaning are not differentiated is too big. This ambiguity could be eliminated with a real Web Thesaurus where meaningful single and compound keywords representing concepts are related by level of specificity to all conceivable Human Knowledge Disciplines. This sounds trivial but it was not easy to perform till now because it means mapping nearly 10 million concepts. Once disciplines and their respective consensus Curricula are precisely defined the whole Web should be "combed" for each discipline proceeding to hierarchically extract for each of its subjects its corresponding keywords set.

While we were investigating how to make a Prototype of this Web Thesaurus evolve by itself through time, we confirmed one crucial conjecture of our Human Knowledge Mapping methodology (see www.intag.org) that states the retrieving power superiority of a Two-dimensional Search by pair [keyword, subject] instead of the One-dimensional search performed by a sequence of unstructured keywords. Definitively unstructured Search Engines should be replaced by structured ones, at least Search Engines structured in two dimensions with the capability to evolve by themselves. In order to prevail, conventional search engines should structure their references like a huge, evolutionary and exhaustive Encyclopedic Virtual Library.

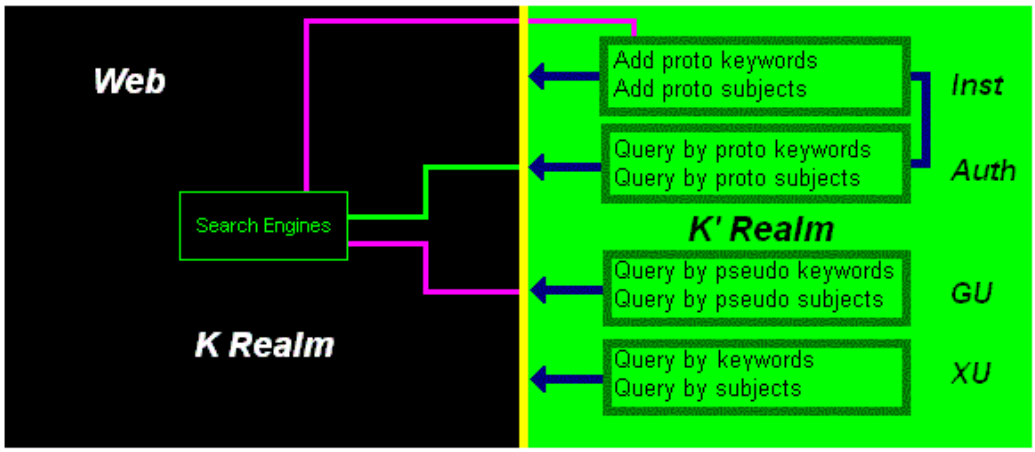
We tested conventional Search Engines twofold: by our Web Thesaurus prototype and by a combination of two logical well known Encyclopedia skeletons: Britannica and Encarta, both in English. The central idea was to search the Web using Conventional Search Engines not directly but through special “semantic lenses” designed to “see” the “hidden order” instead. “Encyclopedia lenses” enable general users to locate what they look for in only one click. Of course this extraordinary efficiency is restricted to a cognitive universe of 80,000 topics. Not bad but not enough!. Our Web Thesaurus may for farther if built to cover the whole cognitive spectrum estimated in 500,000 topics and providing a more precise knowledge skeleton then Britannica and Encarta together adding two levels of detail.

The future of the search industry lies in building and maintaining Web Thesauruses that cover all human knowledge activities, huge logical trees of nearly 200,000 subjects that along five to six levels continuously cover about 500,000 topics, and 10 million of concepts. Next generation Search Engines will have a Universal Wizard that for each language will guide users to find what they need, ideally in only one click of their mouse. All documents should be indexed by these Thesauruses.

The Two Realms of Cyberspace

We have built a Web Thesaurus for a single discipline ICT Information Computing and Telecommunications that has 53,000 keywords related by “level of specificity” to the Association of Computing Machinery 2001 Curriculum. This Thesaurus was automatically extracted from the Web by an information retrieval multi agent. Along our findings we made a clear distinction of two Human Knowledge realms, closely related and complementing each other: the Web itself where the “established” knowledge is represented today by nearly 8,000 million pages, and the “people’s knowledge” that is represented at any moment by all people using the Web. That is crucial if we look for a Web Thesaurus that evolves along the time.

To back up our AI methodology capable to retrieve data and “intelligence” (the hidden order) of the Web and huge reservoirs we needed a new Knowledge Management Paradigm. This paradigm deals with two realms: the K Realm, where Web pages are hosted, and the K’ Realm, where users are connected to the Web space. The paradigm is supported by a set of conjectures that need to be scientifically tested. One of the paradigms tell us that people as “users” meanwhile staying and interacting in the K’ Realm “speak” and even “think” different from people as “Website owners” meanwhile staying and interacting in the K Realm.



For this reason we came to the conclusion that creating the Web Thesaurus will solve the universal searching problem even though restricted to the structured “Established Knowledge Truth”, in fact “half” of the “Human Knowledge Truth”, not bad but half of the ideal and using only half of the Internet’s power.

Our first prototype demonstrates that to have a Web Thesaurus is only a matter of some investment. With a Web Thesaurus Conventional Search Engines could become Super Search Engines by providing what users need in only one query.

To go further we need two Thesauri instead of one, namely: The Web K Realm Thesaurus and the People’s Thesaurus as well. In order to thoroughly test our hypothesis we signed a Research and Development Agreement with the CAECE University from Argentina (www.caece.edu.ar) known for its excellence in Mathematics and Systems Engineering. The tasks initially assigned to the AI-Lab of the university were: “Recognition and analysis of the K’ Realm” and “How people search”.

In the figure above we depict the ways people search in order to reduce their uncertainty:

GU – General Users
By querying with pseudo keywords
By querying with pseudo subjects

This is the most common search. People know what they need but they do not know how to query the established knowledge in the right way. People are by “de facto” enforced to use strings of the established keywords to retrieve what they need. Watching from the K’ side these strings could be considered “pseudo keywords”: potentially K’ keywords.

XP – Expert Users
By querying with the right keyword
By querying with the right subject

This is the search of experts, people who know perfectly well how to look for something in the Web.

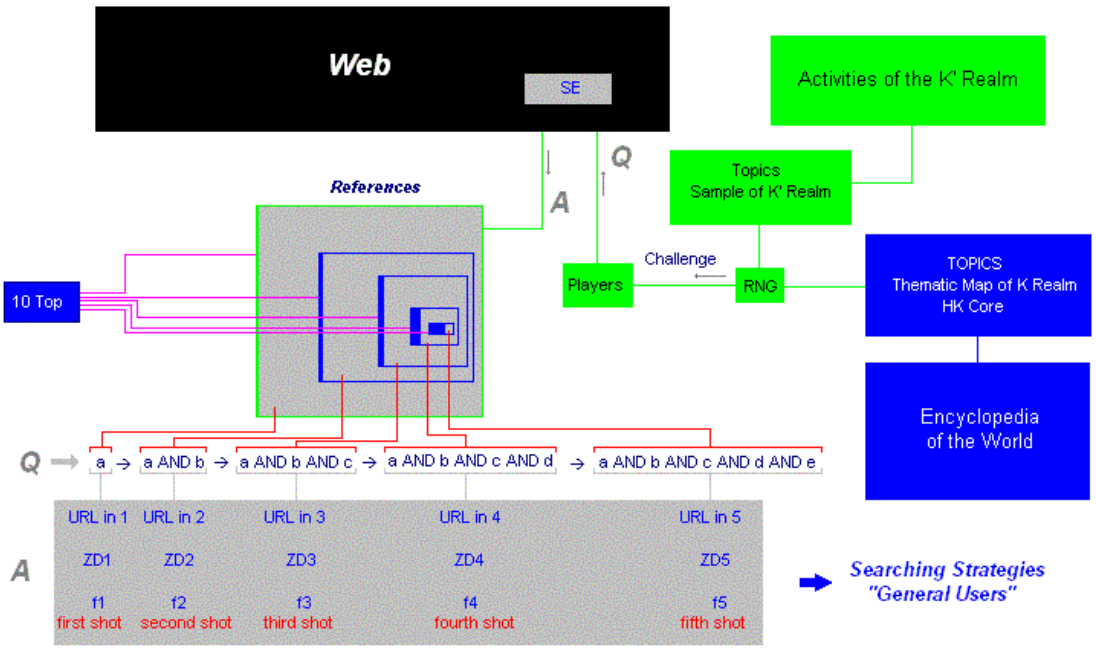
Inst - Institutions
From time to time Institutions like the ACM for computing and the JAMA for medicine proceed to issue standards under the form of new documents with some “neologisms” and with new “subjects”: like for instance “net-centric”.

Auth - Authorities
Once Institutions proceed to issue new standards people that belong to “Authorities” and “Professional Institutions” proceed to query and to use of the new terms intensively. As Institutions and Authorities have a high “popularity” they rank high in the Search Engines outcomes resulting in a sort of feedback that make proto keywords become keywords.

In a near future as long as a People Thesaurus is created frequent pseudo keywords and pseudo subjects should induce new proto keywords adding (fuchsia cycle).

How people search

In order to investigate how people search we envisaged the following experiment. An open and free set of players is challenged to a search championship: to get the truth in as few “shots” as possible when challenged to find valuable information about a given topic. The topic is extracted at random out of a Topics Sample of K’ Realm, where “common people” interacts. The sample is in its turn extracted out of the Activities of the K’ Realm. At the moment the AI-Lab staff defined about 3,000 human activities and their respective knowledge needs.



Players are encouraged to (Q) query the Web via a given Search Engine (or a pool of them), getting (A) answers under the form of “References”, like for instance Google References: paragraphs with an URL and some other indicators of indexed pages. The query game proceeds by “sessions”, with a session consisting of a set of trials and each trial consists of a sequence of up to five “shots”. The valuable answer could appear at any shot. The instructive tell players to only browse the “10 Top”. At any step an agent captures the corresponding “abundance factor” of keyword strings (amount of referred pages) and its corresponding ZD, “Zoom Down” factor is calculated. The trial could “fail” at any step when either the outcome is null or when no valuable URL has been found.

The probability distributions of successful trials track lengths, failures, and ZD’s will be determined as a function of language, players’ knowledge level, demographic and location data.

Outstanding Finding

As the AI-Lab did not have the Topics sample of the K’ Realm ready, we as its Intag counterpart decided to test the people search strategies model with a Thematic Map of the Human Knowledge Core hosted (but hidden) in the K Realm. It seemed to us that good approximations of it were the Britannica Encyclopedia and the Encarta Encyclopedia of Microsoft. With our agent we captured their Logical skeletons –only the skeleton, not the content- and merge them, obtaining about 80,000 subjects organized in a three levels Logical Tree. We provided players with topic, discipline and sub discipline to which it belongs.

Our surprise was that in despite the noise and ambiguities generally present in the Top 10, which by the way was easy to clean by our agent, most searches were successful with only one query!. That confirms one of our milestones: to look for pairs [k, s] keyword, subject, is like adding a second semantic dimension to the searching process!.

We already determined in our first prototype that this second dimension reduces the uncertainty more than 8,000 times in the average!. Concerning the tests performed with Google its Rank algorithm summed up to the second semantic dimension build the miracle!. In despite of noise and ambiguity the algorithm that pumps up authorities and the "semantic skeleton" that enables a second dimension make a strong association to comb the Web efficiently.

The art of creation of meaningful keywords

Let's play a little with keywords. In the ACM Curricula NC means "Net-centric computing", a subject of three single words. Perhaps "centric" was a neologism coined by the ACM Working Teams when dealing with its 2001 ACM Curricula. If we search by [Net AND centric AND computing] we get 13,100 results. If users are not well informed they may try search by [network AND computing] getting 8,440,000 instead!. Subtler users may try a better approach using [net AND centered AND computing] having 182,000 results. We are talking about results, analog to saying the string queried abundance, not about the quality of them, particularly within the Top 10.

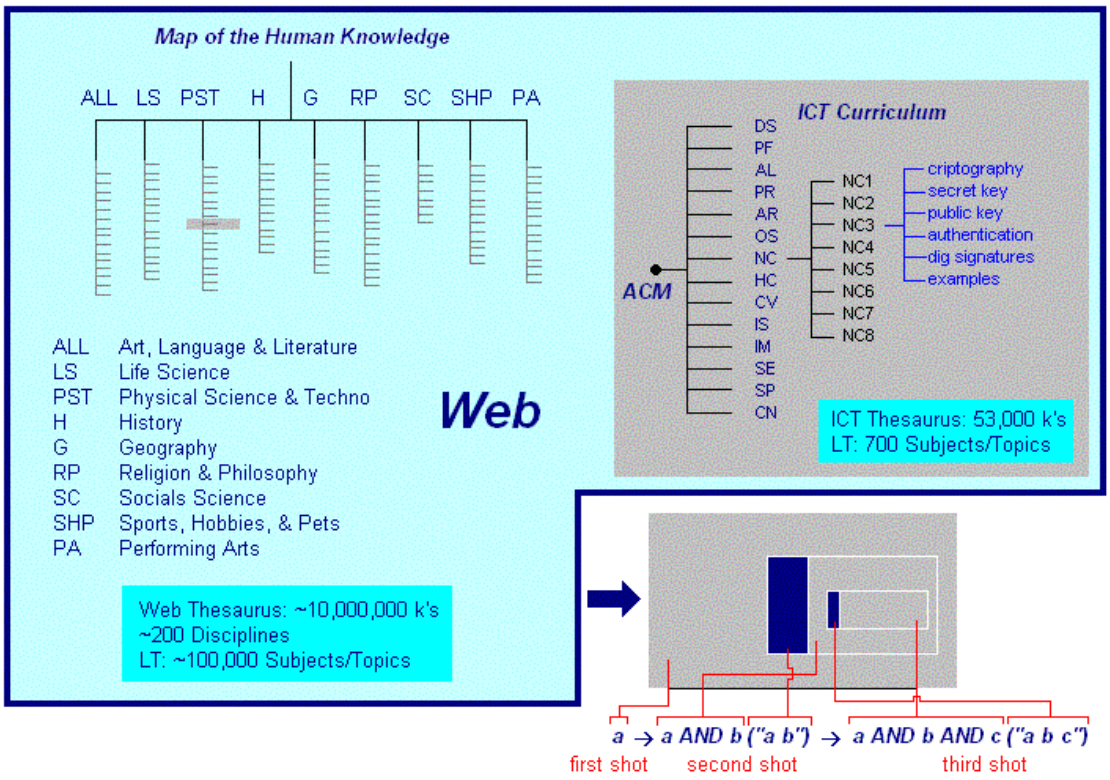
Net-centric computing, 13,100
Network computing, 8,440,000
Net centered computing, 164,000
Net centric computing, 182,000
Net-centric computers, 7,500
"Net-centric computing", 2,640
"Network computing", 1,710,000
"Net centered computing", 15
"Net centric computing", 2,640
"Net-centric computers", 9

Graphics and Visual Computing, 1,190,000
"Graphics and Visual Computing", 373
Graphic design, 8,740,000
"Graphic design", 5720,000
Visual computing, 3,600,000
"Visual computing", 41,000

Next Generation Search Engines

We get to the following conclusion, independently of the search experience as planned with a sample belonging to the K' Realm. The use of a thematic structure like the one provided by the Britannica will significantly improve conventional search engines efficiency. Of course users that want to get answers in only one query will be restricted to a universe of 80,000 subjects and approximately the same number of keywords.

A by far better approach is to build the Web Thesaurus that provides about 120,000 subjects with a semantic universe of 10 million keywords, with an average of 50,000 keywords per major discipline of the Human Knowledge. The ideal approach is then to build the Web Thesaurus but make it evolve as a function of people's interactions.



Nine major disciplines of the Human Knowledge are depicted above. From Physical Science and Technology emerges our first Human Knowledge map prototype for ICT. This logical tree has three meaningful levels. Aided by these glasses any user may go straight to the right target. Before querying by keyword (k) user will be invited to pick one of the disciplines where that k has meaningful acceptations. Once within that discipline a wizard may guide him/her to pick the right target in only one query. Let's suppose that the answer for that query is too large yet, 150 references. A cleaning agent may eliminate and/or kill ambiguous references and another agent could make meaningful clusters with the remaining references. In the figure at the corner below right users and/or agents may experience different alternatives of search, by AND's and by strings

Strategy to precisely retrieve information from the Web

Everybody knows pretty well how difficult it is to retrieve with precision information from the Web. If all pages hosted were indexed via a Web Thesaurus it would be straightforward, most times in only one click. Such a Thesaurus would consist of all existent keywords in a given language associated to a Map of the Human Knowledge. In fact it would be an index of 10 million keywords associated to nearly 200 Human Knowledge disciplines.

Appendix

Details of AI-Lab Search

We've seen that a small alteration of pair [k, s] means to fall into a deep uncertainty hole. Another phenomenon is the intrinsic weakness of actual Encyclopedia when we intend to search specialized information. For instance when searching for

[(acrylic), (Physical Science and Technology => Industry, Mines, and Fuels)]

When we query by acrylic we get in Google 3,830,000 references, and within the 10 Top we find references belonging to the following “disciplines”:

(2) Arts and crafts (working with acrylic)
(3) Painting (with acrylic)
(1) Painting Association
(1) Fiber of
(1) Mother Boards (acrylic in…)
(1) General articles of
(1) Cooling (acrylic parts in …)

That’s a too wide span of ambiguity. This type of ambiguity will be overcome with a Map that adds at least two more level in the logical skeleton.

Another example

"palm reading" 61,300
"palm lines" 819
chirology 3260
"medical chirology" 2
"chirology reading" 0
chirology diseases 101
"chirology diagnosis" 0
palmistry 205,000
"hand analysis" 20,600
handreading 1,990
Dermatoglyphics 4,970
dactylogical reading 50

Where keywords highlighted blue are the “right” ones when searching for references dealing with diseases marks and signs in palms. If you try with the others you may fall in a deep hole of noise and ambiguity. In this case you may appreciate the pressure of the “mode”: originally the “right” term to point to that subject (diseases marks in palms) was “medical chirology” but now it is out of use and the mode goes to palmistry as the right keyword to designate palm reading either for fate prediction or for health diagnosis. Lately the term “hand analysis” was used to make specific professional reference to health and psychological diagnosis. Palm reading could still be used but you may find more abundant and specific references using palmistry.