Kim H. Veltman

New Approaches to Searching, Teaching, Repairing, Advertising and Selling using SUMS (System for Universal Media Searching).

Unpublished, Toronto, 1996

---

1. Introduction

Browsers have become one of the buzzwords of the Internet. The most elementary versions such as Mosaic and Netscape, required that one type in the precise address where one wished to go. A next stage entailed methods such as Yahoo or Lycos which arranged long lists on a given topic. Search engines such as Opentext effectively searched for all occurences of a term on-line. While this produced a great wealth of hits, it made no distinction between different contexts of a word, with the result that many or sometimes even the majority of the references proved to be of no serious interest. All of the above methods are philosophically equivalent because they assume that searching is merely a question of brawn rather than brain; that a better search engine is bigger, faster and gets more hits.

SUMS begins from a very different premise: quality is more important than quantity. Indeed if there is too much noise in the information one receives then quantity can obscure and even undermine quality. How can we find what we really want and not just that which has the same words as that which we want? This applies not only to the initial web surfing when we first find something, but also to how we organize things such that we can find them back more easily thereafter. Indeed the most basic version of SUMS focusses on organizing what we find.

2. Manual Criteria

The most basic version of SUMS is effectively a bucket for collecting information, but differs from simple databases insomuch as it provides a framework for organization in terms of types of access, levels of knowledge, kinds of media etc. The user gathers material manually and organizes the information in terms of six basic questions: who?, what?, where?, when?, how?, why?. Correlations between questions are then established. For example, if who refers to Leonardo da Vinci, links are made to lists of his art, books, buildings, instruments etc. While the initial imput process is tedious, it results in many meaningful connections. Some of these connections are summarized in figure 1.

In the past one would have created a database and used this to correlate various facts concerning Leonardo. The present trends in object-oriented programming mean that these all these parameters can be integrated in a new way. The Leonardo "object" now becomes a series of parameters about the individual with references or pointers to where the materials are kept. Hence, while the complete works of Leonardo may occupy a considerable amount of space even digitally, the Leonardo "object" need only occupy only a tiny fraction of that space. This economy is precisely why it is possible to have a centralized reference base for all persons, objects, places, etc.

Level of knowledge is but one of ten basic organizing parameters. The others are (kinds of) access, (types of) learning, media, quality, quantity, questions, time, space and tools. Each of these can be used to provide finer granularity in terms of organizing and subsequently searching for materials.

Universal (Person)		
Particular (Leonardo da Vinci):	1. Terms (Charts, lists)	Family tree, dates, classifications
	2. Definitions	Biographical Dictionaries
	3. Explanations	Encyclopaedia articles
	4. Titles	Art
		Books- Primary
		- Secondary
		Instruments
	5. Partial Contents	Abstracts
	6. Full Contents.	

Figure1. A summary of some of the concepts associated with the name of Leonardo da Vinci arranged in terms of levels of knowledge.

3. Semi-Manual Searching

A next level of complexity uses a search engine such as Opentext to find materials, but rather than initially searching across the entire web for everything possible, it begins by going to a major classification system such as Library of Congress and searching for the the term therein. This results in a small number of choices. The person decides which sense of the term is most relevant to their needs and then downloads the materials associated with this term. This provides the user with authors, titles etc. relevant to the term in question, some of which will prove relevant in expanding the search. A further stage would successfully download this material automatically into databases under who?, what?, where?, etc.

4. Semi-Automatic Searching

In the above examples the materials collected relate strictly to the horizons of experience of an individual user. A next stage is to make the process a collaborative one whereby two or more experts can combine efforts in developing a resource.

In a further stage this process is generalized. There is now a centralized server on which there exists a standard list of who?, what?, where?, when?, how?, why?. Every time a user finds a term, be it a person, thing, place, time, instruction or cause, it is checked against the master list. If it is new it is automatically added. Every time new parameters of a term are found this is added to the master list and the information itself is downloaded to the user.

The centralized servers are not concerned with becoming repositories of all the existing content but serve rather, as repositories of references, pointers to and indicators of content. This may include simple thumb-nail images of objects. The full versions of images will remain in their home galleries, museums and libraries. In contrast to present day search engines which focus on finding a given term, this repository would contain the clusters of persons, objects, places, times, instructions and causes associated with any given term. All of these would be associated with a term in the way foreseen by object-oriented programming. However, instead of methods such as Taligent, which focussed only on persons (who?), objects (what?) and places (where?), the SUMS approach extends to three further questions: times (when?), which permits an historical dimension, instructions (how?), which extends to the realms of training and repair (see below section six), and causes (why?), which addresses the reasons underlying events.

This leads to a very different paradigm for searching. The user as a client does not choose a term and then attempt, merely by the use of brute force, to arrive at material directly. Rather the client first goes to a centralized server to learn about the parameters associated with the term in question, employs this knowledge to refine further the parameters of their search and then goes out to search for materials on the Internet.

5. Agents

A fully automated method would entail agents. We foresee agents having different roles than those portrayed in the standard approaches today.[1] Agents require a profile of the user as well as access to centralized databanks of the kind described above. This poses important problems of privacy. If all our preferences and searches are known, our weaknesses and strengths could be exploited. One solution would be to make the general statistics available publicly, and keep the particular links to individuals strictly private in ways analogous to banking procedures today. The banks in general know how much money went out and in of a given banking machine, bank, or city, but do not make public who put in what.

The profile would begin with age, level of education and would also identify details of one's profession. The agent would take this profile and go to a centralized databank containing not the contents but simply the references to basic literature and journals as defined by the standard bodies, associations, societies and committees pertaining to each level of education and each profession. Each corpus of literature thus defined would be culled in turn for standard bibliographies and the corresponding names (who?), subjects (what?), places (where?) etc. Any question posed by a user would therefore go through the filter of these different reference sources.

The cumulative experience of libraries and museums will prove essential for the full development of any systematic approach to agents for four reasons. First, classification systems put isolated terms into context. Hence a user searching for a given term, will receive a vocabulary for other possible terms. Second, systems such as Universal Decimal Classification (UDC) have built into them references to broader and narrower terms, and various kinds of relations such as subsumptive, determinative and ordinal. This means that having searched for a term such as chair, the agent can then be prompted to search for narrower terms, such as lounge chair, or broader terms such as furniture without great effort. The potentials for these modalities have already been built into the SUMS system.

The third reason is more complex. Each classification system is actually a subtle reflection of the culture that it classes. Multiple classifications are therefore needed for international searches because that which a user of the Dewey system calls "sociology" might well come under "culture" or some different heading in other systems. A centralized reference base would therefore have a collection of all the major classification systems of the world. Using combinations of search engines and natural language methods a systematic correlation would be made among the equivalent or most closely related terms across all these methods.

A fourth reason extends this principle into the historical dimension where the problem becomes considerably more complex mainly because the meanings of even the most fundamental categories shift enormously. To take an elementary example: a person who searches for "science" today will find titles on physics, chemistry etc. A person searching under the same word (*scientia* in Latin) in a mediaeval catalogue will find titles relating to "knowledge". To find what we now call science prior to the nineteenth century would

require looking under very different terms such as "natural philosophy" (*philosophia naturalis*), "mathematics", "astronomy" or even "geometry".

So a centralized reference base would entail a new kind of etymological dictionary that traces shifts in the parameters of terms. This will require much more than simple study of changing dictionary definitions. A first step will require collating the changing "see also" references over the centuries. A next step requires systematic study of bibliographies and book catalogues. The latest bibliography gives a standard list of names and titles. One then searches through earlier bibliographies and searches for the headings under which these names and titles occur. These findings are arranged chronologically such that one can see how titles which begin in traditional subjects slowly migrate to new subject categories as new fields emerge. Perspective is an interesting example. The earliest bibliographies list perspective treatises under architecture, geometry, surveying, sculpture or even writing. Only gradually does a category of perspective emerge. In addition to bibliographies one needs to extend this same approach to library catalogues and national book catalogues. All of this may sound uncomfortably complex but it needs to be tackled if agents are truly to find serious amounts of information and not just the obvious.

Thus far most work on agents has been done by computer programmers, those trained in artificial intelligence and in knowledge representation. To achieve these goals requires a new kind of centre led by an individual deeply immersed in historical and cultural problems of knowledge, who will work closely with experts in computers, (expert) systems research and knowledge representation. Such a centre should have close connections with a Faculty of Information Science in order to benefit from cumulative experience on classification systems and related themes. The purpose of the centre would not be content but rather to create links for the references of a centralized database which will prove essential for future agents and indeed any serious middleware mediating between the user and sources of content: the great libraries, galleries and museums. In a sense the purpose of the centre will be to extend the approach of industry foundation classes into historical and cultural dimensions.

As the project develops this new approach to objects and foundation classes will be extended to include other seemingly disparate domains, namely, teaching and training; repair and re-engineering and advertising and sales and even philosophy of knowledge.

6. Teaching and Learning

An object-oriented approach to knowledge may start with obvious references to how something is classed, defined, explained, titles to its primary and secondary literature, partial contents in the form of abstracts and even references for the full texts. It will then require systematic references to teaching and training.

Any subject has associated with it a corpus of knowledge. This is typically reflected in standard bibliographies. A single course hardly ever pretends to cover the entire corpus in that field, even though an undergraduate survey course frequently tries to refer to all the major problems in a field, be it psychology, calculus or even western civilization.

Note that we are back to the problem of references which is the theme of the centralized reference base. In this case there are references between the corpus, and its subsets in the form of a course, a textbook, and an exam respectively. Some part of the corpus is in a course; some part of a course is in a textbook; some part of a textbook is in or, as they say, on an exam. The problem with traditional versions of courses, textbooks and exams was that the flow of knowledge went one way and there was no possibility of re-contextualizing the subset. As a result while the expert preparing the course, textbook or exam usually had a good idea of how it related to the category above, the student reading a textbook had little way of knowing whether that course represented five, fifty or ninety five percent of the field that it described. Similarly the exam usually gave no internal evidence concerning the proportion of the textbook, course or field that it covered.

Given the new approach of SUMS these relationships become reciprocal, which means that any bit of knowledge in the spectrum of corpus, course, textbook, exam can be recontextualized. How this would actually be done, will depend on whether the approach is manual or automatic. In the most elementary case the approach is manual. The teacher begins by choosing a subject, say, geometry at the grade twelve level. In this case the corpus might well be defined by the curriculum guidelines and half a dozen standard textbooks. The teacher begins by taking the material in the corpus of curriculum guidelines, practical translations thereof (e.g. the MSSB's guide by Dr. McCudden) and the six standard textbooks, marking these in terms of who?, what?, where?, when?, how? and why?. They then prepare a course with lessons which are a subset thereof. The course and lessons thus have each of their names and terms linked with the basic names (who?), subjects (what?), places (where?), times (when?), instructions (how?) and reasons (why?) in the corpus. The same procedure applies to the test materials. As a result, an individual finding a given name or term in a lesson can find the context of that name or term, using the levels of knowledge to determine the degree of detail into which they wish to enter. Semi-manual, semi-automatic and perhaps even automatic versions of this procedure are possible along the lines outlined above.

This seemingly simple set of connections opens up a whole new range of possibilities in teaching, learning and education generally. It allows a student and/or a teacher to return to the original context of detailed fact somewhere in a course, a lesson, or even in a test (usually after it has been marked). It also allows the user to get some quantitative sense of that context. If a test or exam deals with five problems in calculus, what percentage is that of the problems in calculus in the textbook, in the course and in the field as a whole? Students can therefore reach a more realistic estimate of how much in a given field they really know, which may be quite different from what percentage they obtained on an exam that dealt with a small subset of that field. This information is equally useful for teachers in higher level courses trying to gauge their background knowledge and subsequently to employers who are interested in what their employee can really do as opposed to what mark they received on some exam.

In the past the whole relation of corpus, course, lesson, and test in relation to the individual student was very hierarchical. The ministry of education, faculties and institutes of education, school boards and teachers in a school created a pedagogical great

chain of being that somehow connected the universe of knowledge with the individual student. The hierarchy was largely necessary because the system destroyed the context of the knowledge it conveyed at each level. For instance, the teacher knew (or at least that was the assumption), how the course related to the larger corpus of knowledge on which it was based. The student did not. By contrast, the SUMS approach, precisely because it recontextualizes the individual items of knowledge permits a learning based, student-centred approach to knowledge. Ironically, this approach is so often discussed in the rhetoric of faculties of education whose structure prohibits that to which they say they aspire. Thus SUMS opens new possibilities for learning as well as teaching.

7. Training, Repair and Re-Engineering

In the realm of practice, teaching focusses on training, which is typically about instructions (how?, how to make? how to build?, how to do?). A subset of this corpus deals with a more limited question of how to repair or how to overhaul, now fashionably called re-engineering. Since this need to repair often occurs in unexpected contexts there is a trend to call this Just in Time On-line Learning (JITOL). It will be noted that these problems of training, repair and re-engineering are variants of the referencing problem. Each object, particularly mechanical objects, in addition to their universal concept, need to be linked with their various instantiations in terms of kinds, brands, components, parts and individual examples. Once this becomes part of the reference knowledge concerning an object this is equally applicable for purposes of teaching, training, testing, repairing, and re-engineering.

8. Advertising and Sales

The same principles open new possibilities in advertising and selling products, which is all the more welcome because initial assumptions about the consequences of the Internet for new business have frequently been flawed.[2] Rather than trying to divert the user's attention with an advertisement on a home page while they are in the act of searching for something very different, it is preferable to relate advertising to context. So if a user is searching for campsites under the heading of tourism, then advertising might focus on camping equipment. Ideally information about products, their description, costs, performance ratings, compatibility standards, sales information, market trends of the company could all be available at any time.

At the moment, if I work at a computer all day, rather than hunting on the web to find something I may appreciate a chance to go out to a mall and be sociable. If the Internet option is to be more attractive, it must offer me something the mall alternative does not. If, for instance, there was a guide who helped me decide which products were best suited to my budget, my desires and my needs, this would be very attractive. There might of course be a kiosk version of this at the store in case I decide to go to the store anyway.

If all of these functionalities are combined within a single system, then my criteria for buying can be extended to checking if the product has good on-line instructions

concerning use and repair. Hence, the same system that improves the searching potentials of my agent, transforms my approaches to how I learn, how I repair and how I buy.

9. Static Knowledge and Dynamic Information

Traditionally there was a distinction made between books that were enduring (classics) and literature that was only of interest for the moment (ephemera). In the twentieth century this shifted to a distinction between library books and news. In the past decades this has shifted again to a distinction between static knowledge, as found in libraries, and dynamic information, such as stock market news which changes by the minute. In the past these different types of information were treated differently and seen as being relevant to different audiences. Traditional knowledge was typically free, the latest information cost great amounts of money through subscriptions to on-line services.

The object-oriented approach to knowledge organization outlined above promises to create new links between these two information types, treating them as parts of a single spectrum of knowledge. The stimulus for this trend has come largely from the business environment of the office, where traditionally static records are being connected, through Object Linking and Embedding (OLE), with dynamically changing spreadsheets. In future this principle can be extended to various databases which are changing regularly, such as catalogues of libraries and museums which continue to grow.

This combination of static knowledge with dynamic information will have substantive advantages. Too often those concerned with the latest news have little knowledge of the larger political, cultural and historical context; that the events in Bosnia, for example, relate to Balkan problems that go back for centuries. The combined reference base would allow one to understand these larger contexts. At the same time it would allow a historian to relate past issues to contemporary events and developments.

10. A New Philosophy of Knowledge

Marshall McLuhan once said that the medium was the message. He focussed his attentions on what he perceived to be a shift from print culture to electronic culture, which for him was primarily in terms of television. If we take electronic culture to mean the emerging network based computers which are loosely referred to as the Internet, then it can be argued that the implications of the new media go far deeper than even McLuhan suggested. Involved is a whole new philosophy of knowledge, or rather a transformation in the meanings of both philosophy and knowledge.

Ever since the so called dawn of Western philosophy with Plato and Aristotle, there has been an ongoing debate concerning the relation of universals to particulars. Plato emphasized abstract concepts. Aristotle focussed on concrete particulars. Almost everyone since has either developed these extremes or tried to find some way of moving seamlessly between abstract concepts and concrete particulars. The problem with the latter was largely quantitative. The concept of a chair was simple. A list of all chairs, keeping track of all chairs was impossible for a single person or so it seemed. Given the

developments of AM/FM in the sense of Area Management and Facilities Management especially as it becomes integrated with developments in Geographical Information Systems (GIS) and Global Positioning Systems (GPS), the problem of listing all chairs is no longer an insuperable one. Nor is the idea of being able to track patterns in advertising of chairs, sales and repairs thereof. The case of chairs may seem of limited interest, exciting only to furniture sales personell and accountants responsible for inventory of furniture.

Philosophically, however, even the seemingly dull topic of chairs is fascinating, because it transforms the universal-particular discussion. It is no co-incidence that the best minds in artificial intelligence and knowledge representation have been focussing increasing attention to this set of problems: sometimes in terms of concepts and individuals, which are variously termed instantiations, brands, kinds, components, parts and examples.

If all this is achieved through the centralized reference bases described above it will mean that we can ask whole new sets of questions: not just about where is there a copy of Dürer's book, but where are the known examples of the book, what is the history of their printing, their dissemination, evidence of their influence etc. The scope of things to be known will increase immeasurably. Much of the traditional grunt work of scholarship will be passed to machines, while persons focus on reconstructing partial evidence, analysing trends in the information concerning particulars. Within the course of the next generations, machines will be able to deal with most problems concerning who?, what?, where?, when?, and even how?, such that persons can concentrate on the supremely human question of why?

11. Conclusions.

A central thrust of the new approach outlined in this paper is that a combination of pipeline and content may produce an information highway but cannot produce a knowledge highway with its interesting roads, streets and paths. And while the rhetoric may rightly be in the direction of distributed knowledge systems, a systematic approach to knowledge requires centralized reference bases to act as starting points for those searches through a distributed network of libraries. Otherwise a person with only 100 names in his personal database will have no way of accessing millions of other names systematically. For while they may be able to go to a major library to collect a thousand or even a million names, they will usually have no way of knowing how compatible the naming standards of that library are with those of another. Centralized reference bases will deal comprehensively with the problem of variant names, that a Leonardo da Vinci under Leonardo is the same as Vinci, Leonardo da or even Da Vinci, Leonardo in another catalogue.

Such a centralized reference base thus offers a pragmatic interim solution to the seemingly insuperable problem of standards. There was a time when computer specialists imagined that a day would come when every major collection used exactly the same database with the same headings and the same protocols. They have since learned that even if great institutions had no nationalistic tendencies, it would still be difficult if not

impossible to re-write all the conventions in card catalogues of collections that are centuries old. A centralized reference base will solve this problem. It will also enable agents to become more than electronic caricatures of butlers, transforming them into something useful. Indeed it will make possible many of the high ideals mentioned by proponents of global networks and universal knowledge systems.

A new order is possible with respect to how we approach knowledge, how we access it, how we organize it, and how we analyse the insights that we gain from it. That is the good news. The bad news is that this is not nearly as simple as many have imagined. It is relatively simple to scan in every manuscript in the Vatican. Most persons forget however that these texts are mainly in Latin, with a number in Greek, Arabic, Hebrew, Chinese, and just about every other major language. Most of us do not know that many languages and even those who do would need extensive and intensive training in the reading of old scripts. So after the scanning there will a challenge of translation. Meanwhile we desperately need a centre for the creation of middleware, perhaps a new McLuhan Centre in Culture and Technology, which will prepare the way for making distributed content linked by pipelines sufficiently compatible that the whole will be much greater than the sum of the parts. That was Aristotle's phrase over two millenia ago. If he were writing today he might speak of SUMS renewing this promise of things greater than the sums of their parts.

Perspective Unit, McLuhan Program, University of Toronto
31 January 1996.

Notes

[1] In the popular imagination agents are discussed as electronic butlers who go out and find everything we need. In this model, the agent is active, the user is passive. It is assumed that this can be accomplished using genetic algorithms, which learn as they go. The assumptions underlying this approach are more disturbing than may immediately be apparent. Genetic algorithms entail a transposed version of Darwinian evolution: triumph of the strongest and the most successful. Hence connections which obtain the desired information or which are popular are re-inforced, while "unsuccessful" ones are weakened or eliminated. This means that success becomes a popularity contest, much in the way that classes are tending to become today. Unfortunately, profound thinkers, who may not be very popular because of the difficulty of their ideas, would lose out in this approach. In analogy with cities the biggest would be best, such that New York and Tokyo would be paradigms for living, overlooking the fact that many of the most successful and interesting places to live are smaller towns, villages and the country rather than the biggest cities. We need more emphasis on the paths, streets, roads of knowledge and less emphasis on the information highway. Hence we need a different kind of agent.

[2] One early assumption was that advertising might disappear altogether. In the past advertising was linked with television and since one could track the number of viewers, the most popular times, such as the six o-clock news, the Saturday night feature, the playoff game were prime time and thus more valuable for advertising than, say, a slot at six a.m. on a Sunday morning. It was feared that the advent of 500 channel television

might destroy these simple parameters for quantitative popularity and thus undermine the potentials of advertising.

A second strategy sought to establish new incentives for watching. For example, if a cable channel had pay-per-view movies, then by agreeing to have advertisements in the form of intermissions might give one either a reduced rate on the movie one was watching or some bonus points towards the rental of a further movie.

Internet strategies have focussed on inserting advertisements on the home pages of browsers and the opening pages of indexes of services. Some companies are negotiating their rates for these advertisements in terms of how many times there is actually a hit on the advertiser's logo which takes the user down to the home page. Thought is being given to home pages which lead one deeper and deeper into content and dissuade one from leaving without either making a transaction or at least gleaning some information concerning the web visitor who is a potential customer. There are at least two flaws in this approach. The first is a practical one. If I am searching for something specific, such as medicine or physics, then I will consider any advertisements on home pages along the way as noise, potential distractions from my real purpose. Second, this approach assumes that the primitive state of today's web pages will continue. In fact, once agents are improved to the next level, they will go directly to the subject being searched, thus bypassing any other topics along the way for the same reason that I ignored them as a user. They are not relevant to what I am searching at the moment.