THE CONCEPT OF INFORMATIONAL ECOLOGY
Or
Interest of the information re-use in the company

Luigi Lancieri
France Telecom R&D
Email: Luigi.lancieri@francetelecom.com

Key-words: Internet, caching, user' profiles, Web-mining, information structure, economy,

Abstract: Ecology is an increasing concern of the modern society. In much of fields, one tries to recycle materials. The goal is as much to save transformation energy that the raw material itself. This observation can be transposed and one can also conceive ecology in the field of the information sciences. As in the material field, rough information needs many expensive operations, which can be partially saved by the re-use of data at intermediate stages of transformation. In the context of the Internet network this principle takes a new dimension being given the profusion of data. This document describes various aspects of the metaphor of informational ecology in the context of the communicating company (Internet, Intranet, etc). We develop different methods of initial information structuring as a very important preliminary stage before recycling. Through examples showing the advantages and the problems of the information re-use, we propose methods and variables of control allowing putting in operation a strategy of information recycling.

1. CONTEXT, ADVANTAGES AND PROBLEMS

Ecology is an increasing concern of the modern society. In much of fields, one tries to recycle materials. The goal is as much to save transformation energy that the raw material itself. This observation can be transposed and one can also conceive ecology in the field of the information sciences. As in the material field, rough information needs many expensive operations, which can be partially saved by the re-use of data at intermediate stages of transformation. Thus, the concept of recycling is related to economic problems in the broad sense. The data, that we consume, are processed, refined or summarized. That it is about television, radio, newspapers or now about Web site. The information is prepared to be presented in an attracting way and put at the style of the day. In short, it contains an important value in term of energy and processing time whereas the essence innovation is sometimes minor. It is this latent energy in preprocessed information that we seek to exploit. With the growing success of Internet and corporate networks, these problems took a new dimension. The sharing of information makes it possible to expect a benefit even larger of its re-use. The inter-connected processing systems are as many rich sources of information of the modern company.

In fact, the idea of the re-use in the computer science is not new and can be extended to fields vaster than that of the information. American studies (DOD, 1996) showed that the data-processing code re-use allows a reduction from 20 to 50 % of the development time. This example shows that the benefit of recycling are potentially important even if the problems of the re-use of data-processing code and that of information are not completely the same ones. All in all, the advantages of recycling can be more or less direct.

In term of direct profit, one can highlight that the fact of re-using information already present locally makes it possible to save communications capacities and processing time. This is particularly important in the context of the Internet access because local provisioning of data organized coherently with the company's centers of interests can save important latencies. In this case recycling potentially makes it possible to reduce the obstruction of Web for the benefit of all. In addition, the preliminary standard structuring, necessary to recycling, also makes it possible to built a single interface of access to all the information system of the company. This allows considering the use of standard methods, tools of treatments or generic competences. As many

advantages that carry simplifications effectiveness and economy.

Recycled information can also produce indirect benefit. Indeed, some data circulating in the company are associated with the users' activity and carry their print or their intention, which is very significant from a cognitive point of view. The contents inform us about the users' centers of interest and about their evolutions as well as the interactions between the individuals or their competences for example. In the context of Internet technologies several components can usefully be used for taking advantage of their contents at ends complementary to their traditional uses. We will insist on the interest of the proxy-caches or the News servers, for example. The space of storage of these systems is known as collaborative or symbiotic because it contents result from the interaction between the human and an autonomous memory involving an internal decision system (Lancieri L., 2000 b). Thus, the data contained in these systems can be recycled to carry out new functionalities like metrology and help to the information selectivity. The functions of metrology aim at extracting the print from the users contained in the data and to present it in the form of variables easy to handle and compare. One can measure several types of variables, for examples sets of themes, behavioral or, related to the interactions between the users or, between the users and information. It is possible, for example, to measure the level of informational synergy in the company. This variable measures the coherence of the centers of interests. If the users are interested massively in the same things there is a great synergy. The selectivity of information aims at reducing the space of information implicitly so as to facilitate its exploitation. Indexing the contents of the proxy-caches can do this. It can be hundreds of G bytes of information, which has the advantage of being coherent with the centers of interests of the company. We develop these examples in the continuation of the document.

Although the principle of informational ecology is interesting, it poses several problems. In particular concerning the choice of the type of information to be re-used and for which objective. It is also necessary to wonder about the treatments to carry out so that the result of recycling is useable. As many questions that can call into question the interest of the information re-use. So that, the company can easily use the recycled product it is necessary sometimes to observe rules of productions (a priori structures: XML, RDF, etc.) that can be constraining and limit the spontaneousness and the creativity. Another difficulty is that in certain cases it is more than documents which one uses but resources that cannot be easily co-localized with the users. Certain Web services like the research engines or other resources that point on processes (CGI, PHP, etc.) on distant data are not easily reusable. Thus, it will be sometimes useful to carry out an assessment to estimate if recycling is useful and economically paying in the informational context of the company. We will see that in certain cases, the cost of the re-use is very weak and gives promising advantages.

2. THE STRUCTURING: A PRELIMINARY WITH THE RECYCLING OF INFORMATION

All researches on the information re-use insist on the making of a structure that makes it possible to better describe the data. Scott Ambler (Amber S.W., 1998), for example, identifies 8 types of reusable structures from source data to more organized associations or complex symbolic structures (template, patterns, artifact, etc). In this study, we distinguish two principal approaches. The first that we call the a priori step consists in making tally the data to produce in an architecture which describe them (e.g. SGML, meta tags HTML, CSS, XML, RDF, hytime, MID, ISMID, IETM, HIS, etc.) (W3C Web Site). The second approach said a posteriori consists in taking information in the state and to try to detect structures from the data or to extract some useful information. DataMining or WebMining techniques are located within this framework. The choice between a priori or a posteriori approaches can be the subject of debates but we think that it is hazardous to bet that a structure even if it is the subject of a standard is really universally used a priori. In HTML, which is standardized for a long time, only 15 % of the pages contain the meta tag «keywords » (Agostini P., 2000) that is supposed to give indications on the semantic contents of the Web pages (e.g. for the use of the research engines). In all the cases an important difficulty consists in extracting and exploiting implicit information non-obvious in the data but necessary to their comprehension. These data can be obtained only by taking into account the context or the user profile. In the a priori approach, the stage of creation of information is tiresome for it is necessary to use a minimal formalism even if tools facilitate this operation. On the other hand the re-use is largely simplified since all is described. In term of difficulties it is the opposite for the a posteriori approach, the phase of creation is simplified because without constraints whereas the phase of

exploitation is more delicate. Thus even if the ideal seems to pass by a priori structuring, it is felt well, that it is not easy to put in operation and that it is finally necessary to make cohabit structured and well described information with not structured one. It is besides a strong tendency of research.

Anne-Marie Vercoustre (Verscoustre A.M., 1997; RIO Project-Reuse of Information Object) studied the problem of the structuring and the handling of virtual documents with an aim of information re-using. The objective is to use a language making it possible to create new virtual documents starting from a combination of information coming from existing ones. Various procedures can apply to structured objects or not (basis SQL, OQL (Object Query Language), web pages, etc.). The taxonomic approach of the Guts system (Generic Unified Typing System) of Max Mühlhäuser (Mühlhäuser M. et all, 1998) proposes a form of representation of knowledge contained in the hypermedia documents coming from Web. This approach is based on the transformation of the documents (segments of information, hyper-links, etc.) into networks structures, bound by rules or functions that can be defined by the user. Jan Jannink (Jannink J.) tackles the problem of the detection and manipulation of implicit information present in the data by proposing algebra optimized to solve the semantic inconsistencies and taking into account the context of the analyzed data. This algebra provides a set of operators exploitable with the standard language SQL who allow to express initial information under various angles and to facilitate its reuse.

Other researchers studied to structure the Web huge database in the same way that SQL (e.g. DMQL, WebML, W3QL, WebSQL, WebOQL). In addition, other languages like UnQL (P. Buneman in 96) or Lorel (S.Abiteboul et al. in 97) intended initially for semi-structured documents are also potential languages to structure and exploit Web. The reader interested by a detailed study of these languages can refer to (Zaine O.R., 1999). Taking into account the huge quantity of information the AI techniques were quickly candidates to optimize the visibility of information. Common Lisp Web Server (CL-HTTP HTTP), for example, offers a Web interface for AI systems in general and more particularly those based on traditional Lisp. This free distributed product was designed by the MIT within the framework of IIIP project. Objects technologies as Corba recently took the Internet turn (protocol IIOP). One can thus find in freeware the libraries allowing interfacing objects environments with the Web, it is the case of ANSA package (Ansa, 2000). LOOM System (LOOM Web Site) developed within a university framework is supporting the representation of knowledge in declaratory mode. LOOM makes it possible to develop intelligent «processes» exploiting the Web. In term of interfaces, one will note also the appearance of a software component integrating an inference engine making it possible to communicate with the navigators via RDF format.

3. EXAMPLES OF APPLICATIONS

This state of the art shows us that the techniques aiming at structuring the heterogeneous contents do not miss even if research is still active on the subject. It is important to know «how » to structure but it is as much, if not more, important to know «what » to structure. Indeed, one can deal with data that do not have any interest for the company. We describe now some methods allowing to have this essential information with a low processing cost. The interactive shared bookmark (BPI) (Lancieri L., 1997, 2000 b), for example is situated halfway between the search engine and the bookmark. The principle consists in using the data contained in the proxy-cache of the company to constitute a database usable to facilitate the information search. This database would constitute to some extent « a super bookmark » since the content of the cache is a superset of resources contained in the bookmark for all users. One of the interesting points of this system is its adaptability. Indeed, the content of the cache will evolves and follows the users, so it constitutes an information base reduced on the dominant interests of the company. To validate the interest of this system we asked 7 users to make ambiguous requests (e.g. Network) and to give an appreciation value from 0 to 5 reflecting the level of interest for the first 20 results provided by BPI compared to HotBot Lycos (as an example of regular search engine). The average notes are the followings: BPI: 3.2, standard deviation = 1; Hotbot: 0.7, standard deviation = 1. One rather quickly notes some differences between the results provided by both systems. The size of the files corresponding to URLs returned by BPI are big sized textual files compared with those of Hot Bot (128,2 KB on average against 13,4 KB, 10 times more. Consider that a 128 kb file contain around 50 textual A4 pages). It is very clear that the answers provided by BPI are centered on the Internet techniques which are the field of competence of our site (55 % of the answers are 1.7 Mo on the whole). In comparison, the firsts answers of Hot Bot are relatively etherogeneous and strongly trade and publicity directed with very short contents (65 % of the answers are 97 KB on the whole). The first relevant response compared to our centers of

interests provided by HotBot is to the 12-nd position whereas as soon the first position BPI provides documents with strong and interesting technical contents. One will note also the difference of the number of results provided by the two engines. BPI give 300 URLs whereas HotBot provides 50 000.

BPI combines explicit and implicit collaboration of the user by exploiting existing components (proxy-cache) and a generic interface (Web browser + CGI). This approach that limits the specific developments makes it possible to reduce the costs of operation and evolution of the product. In term of effectiveness, BPI makes it possible to associate the power of a search engine with the smoothness of a bookmark. The results often integrate recent sites that are not yet indexed by the traditional engines (e.g. case of the sites transmitted of mouth to ears). We also know that only around 10 % of the Web are indexed by search engines (Lawrence S. et all). In addition since BPI is conceived to collaborate with a traditional engine, it transmits the best «sites »as well as possible (selected according to the local interests, popularity, etc.) and in the worst case the same sites as a regular engine. With regard to the effectiveness, we saw with the users' tests that the implicit filtering used by BPI is effective since research relates to a subject of interest of the group. One would critics this demonstration saying that one would have obtained comparable results by specifying precisely the request in Hot Bot (e.g. tutorial network IP). This remark is partially true but even if one avoids the sites that have nothing to do with the subject (e.g neural networks) it remains much of noises (short pages with little informational contents). Actually one knows that the principal problem of the search for information is to well formulate its request, for this reason the beginners have difficulties to find what they seek on the network. In this context, BPI offers more than comfort while avoiding formulating requests tiresome to express.

The second example of recycling opportunity is the one of News Server. This technology allows to re-uses the knowledge exchanged by the users. The corresponding data contains the complex result of various interactions and have a strong potential. As in the case of social groups, the members of the electronic groups are more or less influenced, often unconsciously by the exchanges of information that are carried out in the group. Some members send more messages than others or some messages are the object of more discussions than others. In some groups the main part of the messages are sent by a minority of members, -the others being more passive-, and thus have a rather monolithic form of influence. On the contrary some groups are very heterogeneous and everyone takes part «in all the directions ». Taking into account the participations of individuals in several groups can highlight these forms of influence.

Our study is based on the analysis of the messages headings posted in 640 Usenet news groups during 2 weeks. Our hypothesis is that the level of influence in a group can be evaluated by the analysis of the distribution of the number of messages sent by the transmitters (see Lancieri L. 2000 for a discussion on this assumption). In particular, we underline characteristics of self-similarity of this distribution. We define the coefficient of complexity of the interaction (CCI) by the value of the slope expressed by the power value D of the Zipf law. This coefficient translates the decreasing speed of the distribution of the number of messages ranked by frequencies. In accordance with the Zipf law the logarithmic transformation curve of this distribution is a D slope line. This line expresses the property of self-similarity that expresses that the increase in popularity is constant whatever the order of increase (Mandelbrot B. 1997). The slope tends to 0 if the messages are equally distributed on all transmitters and tends to infinity if the majority of messages come from limited transmitters. As we say, previously, this coefficient is related to the level of influence within the group. We reproduced in the following figure two slopes corresponding to the "alt.os.Linux" and «intel.inbusiness » news groups that are respectively of –0,6 and –2.5. These values indicate that the messages in the first group are relatively well distributed on all the transmitters whereas they are concentrated on few transmitters within the second group. These results are not surprising because the "intel.inbusiness" group would be much more directed than the freer Linux group. One of the aspects that we did not treat here but who would deserve to be developed relates to the temporal evolution of the CCI coefficient. The variation of this evolution would make it possible to account to us for the evolution of the influence within the groups.

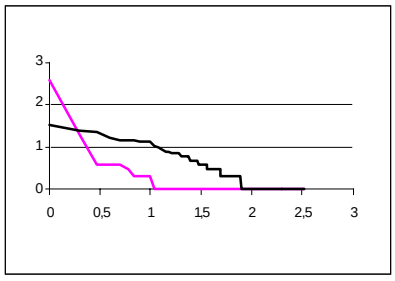

Figure 1: *CCI for the intel.inbusiness «groups »and «alt.os.linux »*

In an academic context, several works focused on the recovery of information based on the interactions between people in the Web has carried out. J.S Donath developed a tool (Donath J., 1995) making it possible to know and contact in real time peoples who consult the same Web pages, thus working out new forms of interactions. The news groups were also subject of many other works. H Sorensen developed a tool named PSUN that makes it possible to extract the thematic profiles from the users according to the articles that they posted. These profiles can be used to recover and filter the most interesting articles for these users. A. Jennings (Jennings A., et all, 1992) used the neural networks to learn the profile from the user and to assist him in his research. With regard to filtering and the re-use of the news groups, one will read the article of F Kilander (Kilander F.,1996) that compared a dozen tools and makes it possible to have a good idea of what was made in this field. The objective of the system suggested by Ygal Arens (Arens Y., et all, 1994) is to preserve the users' requests and its results provided by a search engine. The idea is to be able to re-use this information for future requests. All the answers are not saved only are those that can be represented in standard KL-one semantic system of representation. This method thus implies a decision of storage on a priori criteria. The criteria are in particular, the low size of the answer, the good semantic coherence between question and answer and the representativeness in the existing semantic structure at the time of storage. In other words one does not store the print of the user but the part of this print that satisfies the criteria described a priori. A close system is used in Amazon.com to make bringing together between the customers and to propose books likely to interest them.

4. ESTIMATION OF THE LEVEL OF INFORMATION RE-USE

It is useful to estimate the level of information re-use in the company at least to decide if it is advantageous or not to put in operation some recycling systems. The rate of the Web requests redundancy is a good measurement. This rate, computed from the proxy-caches log files, measures the ratio between the number of the renewed requests and the total number of requests. This kind of files also exists in the Web servers but it is less interesting because the accesses relate only to the servers limited contents. The proxy-cache log files on the other hand relate to more open contents, which are potentially that of whole Web. The hit rate used to characterize the cache performances in its context of use is also a good estimate of the level of information re-use. A high hit rate indicates that the users' centers of interests are very homogeneous and that many documents are re-used. Consequently, many documents are delivered directly by the cache, which accelerates the delivery of information since it is local. The rate of re-use is variable on observation period and depends also on users. It is possible to extract from the users' access the relationship (DA) between the number of different URLs (D) and the total number of consulted URLs (T).

$$DA = \frac{D}{T}$$

This ratio indicates the access dispersion and gives an indication on *the heterogeneity of the users' centers of interests*. More this value is weak (high accesses on the same documents) more user will be likely to see its request been directly sent by the cache (from where weak latency time).

Users	T	D	DA %	
234.68	3582	340	9,490	← High reuse rate
234.98	4138	2928	70,760	
129.8	4269	3190	74,720	
234.27	4147	2654	64,000	
234.71	2658	1902	71,560	
234.208	1800	972	54,000	
234.63	5649	2175	38,500	
234.212	3369	2552	75,750	
129.82	1287	935	72,650	← Low reuse rate
Together	**30899**	**17648**	**57,120**	

Figure 2: *Characteristics expressing the level of information re-use for various users*

The previous table gives, for 9 users and a total of a little more than 30 000 accesses, the characteristics that we evoked. Although this sample has a low statistical representativeness from the point of view of the number of users, we can have some references. To be rigorous, it should be said that these values are only indicative because approximately 30 % of the Web documents are dynamic ones. In addition, part of the dynamic documents, difficult to evaluate precisely but about 15 %, correspond to documents referred by a single URL but which correspond to different contents one day to the other (e.g. stock exchange price, weather, etc). It thus should be considered that the values of the table inform us on the level of information re-use with a tolerance of about 15 %.

It is clear that a company that would obtain a high total ratio would have little interest to operate

an expensive recycling strategy. The investments would be not easily paying taking into account the fact that these centers of interests are too much dispersed.

5. CONCLUSION

As we saw the re-use of data in the company is of several interests! Above all, they are rich information on the behavior or the users' centers of interests. Then these data are low cost to obtain because they are already present in the company and already used for their informative first function. The re-use or the recycling of these data is connected with a form of informational ecology. It contributes to have a unified view of the company information system and helps to economize expensive resources. The explosion of the quantity of data can produce a deficiency in term of processing capacity and an information overload (too much information kills information), the re-use of already processed data can be very useful.

REFERENCES

Agostini, P, 2000 ; France Telecom study on the use of MetaTags keywords in web pages

Abler S.W., 1998; The various types of object oriented reuse; http://www.findarticles.com

Ansa 2000; Web site; http://www.ansa.co.uk /ANSA/ ISF/ wwwCorba.html

Arens Y., C.A. Knoblock, 1994; Intelligent Caching, Selecting, Representing and reusing data in an information server; In proceedings of 3 rd international conference on information and knowledge management; Gaithersburg 1994

DOD, 1996; software reuse initiative; Web site: http://dii-sw.ncr.disa.mil/reuseic/pol-hist/primer

Donath J., 1995; sociable information space; IEEE workshop on community networking; June 1995

Jannink J.; rethinking Reuse information; Web Site: http://www.dyncorp-is.com/darpa/meetings/gradmeet 98/Withepapers/stanford.html

Jennings A., Higuchi H., 1992; A personal new service based on user model neural network ; Telecom Australia 1992.

Kilander F., 1996; A brief comparison of News Filtering Software; Stockholm University; June 96

Lancieri L., 1997; Interactive Shared Bookmark; In proceedings of WebNet 97 -Association for the Advancement of Computing in Education (AACE)

Lancieri L., 2000; A connectionist approach for evaluating the complexity of interaction in the World Wide Web. The case of News Groups. IJCNN 2000 Como (Italy).

Lancieri L., 2000 b; Memory an forgetfulness: 2 complementary mechanisms to describe Internet users in their interactions. PhD Thesis university of Caen.

Lawrence S., et all; Accessibility and Distribution of Information on the Web; Steve Lawrence and Lee Giles http://www.wwwmetrics.com/

Loom; Web Site. Web Site: http://www.isi.edu/ LOOM/LOOM-HOME.html

Mendelbrot, B., 1997; Les objects fractals; edition Flamarion réedition 1997

Mühlhäuser M., et all, 1998; Mühlhäuser M., Ralf Hauber, Theodorich Kopetzky; Typing concepts for the Web as a basis for Re-use; in proceedings of Workshop on Reuse of Web information, The Seventh International World Wide Web Conference Brisbane Australia 1998

Plewczynski D., 1997; Landau theory of social Clustering; Institute of Social study Warsaw November 1997.

Vercoustre A.M. et all, 1997 Vercoustre A.M. and François Paradis, A Descriptive Language for Information Object Reuse through Virtual Documents, in 4th International Conference on Object-Oriented Information Systems (OOIS'97), Brisbane, Australia, 10-12 November, 1997.

W3C Web site; Site Web du World Wide Web Consortium; http://www.w3c.org.

Zaïne O.R., 1999; Ressource and knowledge discovery from the Internet and multimedia repositories, PhD Thesis, Simon Fraser University 1999, Canada.