Semantics and Experience in the Future Web

Enric Plaza

IIIA, Artificial Intelligence Research Institute
CSIC, Spanish Council for Scientific Research
Campus UAB, Bellaterra, Catalonia (Spain)
enric@iiia.csic.es

Abstract. The Web is a vibrant environment for innovation in computer science, AI, and social interaction; these innovations come in such great number and speed that it is unlikely to follow them. This paper will focus on some emerging aspects on the web that are an opportunity and challenge for Case-based Reasoning, specifically the large amount of *experiences* that individual people share in the Web. The talk will try to characterize this experiences, these bits of practical knowledge that go from simple but practical facts to complex problem solving descriptions. Then, I'll focus on how CBR ideas could be brought to bear in sharing and reusing this experiential knowledge, and finally on the challenging issues that have to be addressed for that purpose.

1 Introduction

The Web is a vibrant environment for innovation in computer science, AI, and social interaction; these innovations come in such great number and speed that it is unlikely to follow them. This paper will focus on some emerging aspects on the web that are an opportunity and challenge for Case-based Reasoning, specifically the large amount of *experiences* that individual people share in the Web. These experiences, ranging from client reports on hotels they have visited to small explanations on how to do certain things, are searched for and reused by thousands of people. These experiences can be found in forums and blogs, in normal web pages and in specialized services like Question-Answer websites.

However, they are treated *documents*, not as experiences. That is to say, they are represented, organized, analyzed, and retrieved as any other document. The main purpose of this paper is to argue that there is a special kind of content, namely experiences, that provides a specific form of knowledge, experiential knowledge, and they should be represented, organized, analyzed, and retrieved in accordance to this nature. Moreover, the paper will provide some food for thought by proposing some ideas on the conditions required and the techniques suitable to build systems capable of reusing experiential knowledge provided by other people in specific domains.

The structure of this paper is as follows. Sections 2 and 3 discuss two of the most noteworthy components of current debate on the web, namely adding a semantic substrate to the web (e.g. the semantic web, folksonomies) and the

K.-D. Althoff et al. (Eds.): ECCBR 2008, LNAI 5239, pp. 44–58, 2008.

phenomenon of social networking. Then Section 4 discusses the nature of experiential knowledge, while Section 5 elaborates the conditions for reusing other people's experiences.Next Section 6 discusses the relationship of semantics and experience, Section 7 presents several forms of experience and discusses their properties, and Section 8 proposes a process model for systems reusing experiential knowledge on the web.

2 Semantics, Up and Down

In this section, I want to examine two approaches to imbue semantics in the web content: the top-down approach of the semantic web and the bottom-up approach of social networks. The Semantic Web (SW) [1] was proposed with the purpose of allowing the human-produced web content to be understood by automatic systems: ontologies define the terminology that "agents" use while roaming the web pages entered by humans using SW-enabled tools. This proposal is a top-down approach to semantics, in the sense that someone designs and maintains the definition of an ontology for a given domain. In a new paper revisiting the SW [2] this vision is refined: ontologies "must be developed, managed aged, and endorsed by committed practice communities." I think the conditions are even more restrictive: an ontology only makes sense for a domain if used by a *community of practice* — not just any community that endorses a particular ontology specification. A community of practice (CoP) is developed by a process of social learning of a group of people with common goals, while they interact with the purpose of achieving those same goals. Knowledge Management (KM), initially focused on *explicit knowledge*, has used the concept of CoP to address *tacit knowledge* which cannot easily be captured, codified and stored. From this perspective on semantics, SW and KM share a great deal of challenging issues.

Folksonomies, the bottom-up approach to web semantics, originates from the tagging processes in software platforms for social networks, sometimes called "Web 2.0". Folksonomies are lightweight shallow ontologies that emerge in specific community of practice where users "tag" some content objects (like photos in Flickr.com) with whatever keyword they deem more appropriate. Folksonomies are interesting in that they emerge from the social learning process of a community of practice: people learn to use other people's tags and introduce their own that, if found useful, will be used by the community at large. For this reason, folksonomies are a way to capture part of the elusive *tacit knowledge* in a network of practice (the name given to a community of practice in a social network software platform).

Some people would object considering a bag of keywords or tags an ontology, insisting it is merely a type of meta-data, but so are ontologies. The argument usually focus on the fact that ontologies are structured and folksonomies unstructured, but the main difference is in the way semantics are assigned: while ontologies are based on explicit specification of terms, folksonomies rely on a statistical analysis of the usage of terms in the context of a network of practice. From the standpoint of the philosophy of language, ontologies purport a *logicist*

approach to the meaning of terms: a term is an instance of a concept if and only if it satisfies the concept's definition. On the other hand, folksonomies seem closer to Wittgenstein's notion of *language-game*: a term has a specific meaning by the way it is used in a particular context [3].

Some researchers will inevitably try an hybrid approach combining a top-down ontological approach with a bottom-up user-driven open-ended folksonomy: an ontology may define the explicit preexisting knowledge in a domain while the folksonomy captures part of the explicit and tacit knowledge of a network of practice. Although bridging the gap between both approaches is an interesting research issue, this is beyond the scope of this paper. For the purposes of this paper, the important point is that ontologies, the SW, and web semantics in general, are a *enabling technologies*: a substrate that provides some service required by more complex tasks — not a way to do more complex tasks. Specially the SW seems now to be a platform to develop a specific type of applications called ontology-based systems [4]. At the end of the day, the developers of a new web-based system will have to decide what kind of semantic model is suited for the specific web content they have to work with. The suitability of semantic models to different application domains and type of content is an empirical one, and the future web-based systems will explore and ascertain their advantages and shortcomings.

Let us now examine the existing, most burgeoning new systems in the web: social networking software.

3 The Network Is the Content (or Vice Versa)

"The network is the computer" claimed J. B. Gage of Sun Microsystems to emphasize the importance of network access for modern computing systems; nevertheless, Oracle's "network computer" (a diskless desktop computer promoted by Sun and Oracle) was not a successful answer to that claim. The myriad new software platforms for social networking seem to make a similar claim: the social network is the most important part of the so called Web 2.0. Indeed, the network effect in the web has impressing performance, from Google's page ranking based on hyperlink connectivity to Facebook or MySpace social networking websites. However, social networking is part of the picture but it is not the whole picture: some systems like LinkedIn focus the network of social relationships, while others like Flickr the (photographic) content is the most important part and the social network (as such) plays a lesser role.

From my point of view, what is most relevant is the *user-contributed content*, be it photographs or links to other people: the personal relationships that constitute social networks are part of the content contributed by users. This does not deny that the social networking plays an important role in facilitating the contribution of content by the users, quite the contrary: social networks create wealth and can originate a "social production mode" (see for instance Yochai Benkler's *The Wealth of Networks* [5], that presents a comprehensive social theory of the Internet and the networked information economy). Thus, networking

facilitates the creation/contribution of content, and it is indispensable; but as a social mode of production[1] is a means to an end, namely what is produced: the *user-contributed content*.

Be that as it may, the bootstrapping of social networks and social production of content is outstanding. In this paper I want to focus on a particular kind of content that can easily be contributed by people: their own experience in some domain or other.

4 The Case for Experience

Before proceeding on to discuss user-contributed experiential knowledge on the web we need first to elucidate what the term *experience* means. Case-based reasoning (CBR) may be understood, first and foremost, as learning to solve problems (or take decisions) from past experience. More specifically, past experience is represented in the form of a collection of *cases*, where a case ($situation1$, $outcome1$) is to be understood as knowing that in the past, when what is described in *situation1* held, then the *outcome1* (that may be a consequence or a decision) also happened. Thus, a case is a statement (at some level of description) of a fact observed or experienced in the world. Additionally, CBR systems use case-based inference (also called analogy and similarity-based inference) based on the assumption that when a new *situation2* is similar to an old *situation1* then we can plausibly predict an *outcome2* similar to *outcome1* is correct.

The representation of cases, situations and outcomes may be very different across domains (from k-NN classification to case-based planning); but they have in common that they present the knowledge of an observed factual situation: e.g. "this is a good hotel because my stay was very agreeable", or "I did this sequence of actions (this plan), in this situation, and I achieved that goal". Although there are no "cases" as such on the web there is a huge amount of this kind of *practical knowledge* present today in the web. This kind of practical knowledge coming from the direct observation or experiences of people is what we will call *experience*.

In all likelihood, experiential content in the web is one of the most valuable web resources: people constantly use these resources to decide issues (e.g. booking a hotel, visiting or not some tourist spot) or solving problems (e.g. browsing through a forum on digital photography to learn how to solve some issue they encountered in a photo they made). In economic terms, experiential content is one of the most added-value resources on the web today, and if properly marshaled could provide attractive added-value services.

The technological challenge is how to represent, organize, and reuse experiential content. I surmise that the first step to address this challenge is to recognize that there is such a thing as "experiential content," and not merely hyperlinked texts. The way content is organized nowadays is a network of documents, and

[1] *Social Production* is production of information, knowledge and culture that is not based on price signals or on command structures [5]. Computers are the main means of production and networks those of distribution.

possibly in the next future, *annotated documents* (using ontology-defined concepts or folksonomy-based tags).

Moreover, the way users work with web content is what I'll call *Search & Browse* (S&B). The web users typically need to *first* use a search engine to find a "resource," this may be an external search engine (e.g. Google or Yahoo to find a website or a page) or an internal search engine (e.g. search inside a forum for the posts that may talk about the topic of interest). *Next*, the users need to browse a (sometimes disturbingly) large collection of "found items," perform a cursory read of them to filter out those blatantly irrelevant, then read carefully the rest (while eliminating those subtly irrelevant) to isolate the relevant content. Finally, the users have to *reuse* the relevant content, that may be dispersed in a dozens of pages in different websites; notice that there is no support for the users' task and they simply use "copy & paste" to aggregate the information found or print all those pages and then aggregate that information.

4.1 Found and Lost

A specific example may be useful to illustrate this scenario. Let us consider the task of deciding which hotel to book and consider the existing experiential content of previous hotel clients that describe their good and bad experiences in those hotels. Let us say there are H hotels in the intended destination, W websites with hotel-related experiential content, and each hotel in each websites has on average C client reports: a user to be well informed would need to search & browse $H \times W \times C$ user-contributed experience items. This is a huge amount of valuable information but ineffective if it is to be manually processed, as is the case now in the S&B paradigm where there is no support for the task the users want to carry out, and for which reason they have performed a search in the first place.

Certainly, the users are capable of cutting down the work by filtering out information: by selecting a few websites (equivalent to performing a sampling operation $w = sample(W)$), the reducing the eligible hotels by some hard constraints like "3- or 4-star hotels only" (a filtering operation $h = filter(H)$), and finally accessing a subset of all client reports (a sampling operation $c = sample(C)$), the workload is reduced to examining $h \times w \times c$ client reports. Notice that there is no computer support to perform a good sampling of websites or client reports: the users have no way to know if the acquire a *good sample* of the population — simply having this kind of support automated would improve both user workload and output quality.

Moreover, the real task for the users starts now and is also unsupported: they have to aggregate for each hotel in h a number of around $w \times c$ client reports, e.g. determining pros and cons for each hotel according to the majority opinion of those reports, and finally deciding on the hotel that better fit their interests. Clearly, the S&B paradigm does not support this process and the users end up making a less informed decision. However, AI techniques could be used to support this decision, and I'm not referring to data mining or recommender systems, but to a reinterpretation of Case-based Reasoning that would allow us to support users in using experiential knowledge provided by a community of practice.

5 Reusing Other People's Experiences

Considering again the hotel selection example, we can easily substitute the *Search & Browse* process by *Retrieve & Reuse* processes of CBR:

1. the Retrieve process searches for client reports of hotels close to the declared interests of a user and selects a subset of them; then
2. the Reuse process analyzes the retrieved client reports in order to aggregate the information about pros and cons of each hotel and finally produces a ranking of hotels taking into account both the user's interests and the pros and cons of each hotel.

This mapping is sound, in the sense that both Retrieve and Reuse processes follow the ideas in [6]:

1. given a problem (a specific task to be achieved) the Retrieve process selects the subset of cases (experiential knowledge) most similar (or relevant) to that problem, while
2. the Reuse process combines, in some specific way, the (experiential) content of those retrieved cases (and possibly using some domain-specific knowledge as well) in order to achieve a solution for that problem (that specific task to be achieved).

This rather abstract mapping allows us to determine in what a CBR approach to support experiential reuse in the web add to the S&B paradigm: the definition of a user-defined task to be achieved. Indeed, only when a problem (a specific task to be achieved) is posited then a *Retrieve & Reuse* approach can be used. Let us return to the hotel selection example again. Clearly the kind of hotel the user is interested in depends on the type of travel. For instance, whether it's in a one-night business trip or a week of leisure, the pro and con factors that are important may vary for one kind of travel to another. For instance, the factor of whether the hotel staff is categorized as friendly (in pros) or unfriendly (in cons) depends on the trip: a friendly/unfriendly staff is not important in a a one-night business trip while is quite important on a leisure week travel. This correspondence between the hotel client reports and the user interests would be performed inside the Reuse process, e.g. preferring those hotels with a clear majority of client reports stating a friendly stuff and the other factors important for the user. Notice that this is precisely the work the human user has to do, without any support, while examining $h \times w \times c$ client reports.

Nevertheless, there are differences from the traditional CBR approach with respect to a *Retrieve & Reuse* approach to use the experiential knowledge of other people. These differences stem from tacit hypotheses used in CBR or implicit assumptions built from practice in building CBR systems. A first implicit assumption is that the Retrieve process will select one case (or a small number of cases) on which the Reuse process will work upon. As the hotel scenario shows, this is not the best option when dealing with experiential knowledge coming from a (potentially large) number of people. In the hotel scenario the role of the Reuse

process is to select, among a huge number of client reports, a sufficient number of reports about hotels that are relevant for the specific request of a user (here seems more appropriate to call a user-defined query a *task* or a *request* than a *problem*).

Since the Reuse process needs to aggregate information from disparate sources in order to avoid noisy data, the sample of data has to be large enough so that aggregation methods like averages or weighted averages are meaningful. That is to say, in the hotel scenario the role of the Retrieve process may be to select the hotels relevant for the task at hand within some given ranges, for instance, of price and location, and then gathering all their relevant client reports. Additionally, the Retrieve process could perform an additional filtering or client reports based on their age, client reputation, etc. Then, given this sizeable sample of people's reports on their experiences, the Reuse process may be able to aggregate, from the evidence of disparate sources, the likelihood that one or a few hotels are the most adequate for the particular interests of a user travel.

The robustness of using experiential knowledge originating from multiple sources has been studied in several scientific fields. In Machine Learning, the "ensemble effect" states that using an ensemble of learning systems reduces always the error when compared to any single learning system. The only requirements for the "ensemble effect" to take place is that the prediction of individual learning systems is better than random and that their errors are not correlated with one another [7]. Similar properties have been characterized in Social Choice Theory, where the Condorcet Jury Theorem provides a similar property for taking average measures like voting in a jury [8]. Communities of practice on the web have been known to show a similar effect, a fact popularized in James Surowiecki's book *The Wisdom of the Crowds* — where similar conditions are prescribed in order to insure the emerging effect of wise decision or prediction by aggregating information from a crowd of people.

Therefore, a challenge for applying an approach like the *Retrieve & Reuse* one sketched here is to enlarge the core ideas of CBR, namely reasoning and learning from past experience, to a scenario where experiential knowledge originates from multiple individual sources; this multiplicity would require that we incorporate aggregation measures that obtain the desired "ensemble effect" into the Retrieve & Reuse processes. There are other CBR assumptions that need to be challenged to develop systems that reuse experiential knowledge on the web, and we will summarily address them in the next sections.

6 Semantics and Experience

In this section I will address to more challenging issues that need to be addressed in order to reuse experiential knowledge on the web, namely the semantics and structure of experiential knowledge. Concerning semantics, we have already discussed in section 2 the top-down approach of the semantic web and the bottom-up approach of folksonomies. Both approaches are suitable to be

used in a CBR-like approach to reasoning from experiential knowledge on the web:

1. the semantic web uses ontologies expressed in description logics (specifically the OWL language[2]), which is compatible with the research line on knowledge-intensive CBR systems development using description logics;
2. Textual CBR [9] has been working on a bottom-up and hybrid approaches to semantics in cases expressed as text, which is compatible with the current research goals of folksonomies and web text mining — I think that the natural extension of Textual CBR is to address the challenges of textual experiential knowledge on the web.

Since both semantic approaches, or a combination of top-down and bottom-up approaches, are suitable for a CBR-like approach to reuse web experiential knowledge, the challenges are simply the same of any other web-based system developed using Artificial Intelligence techniques. Moreover, since the applicability and utility of either semantic approach may vary for different application domains, it is an empirical issue to determine when and how these semantic approaches will be useful. In this sense, the approach to reuse web experiential knowledge I'm sketching here would be neutral on these semantic debates, trying to find a suitable trade-off for a particular application domain and to keep up with the new developments in web semantics.

Nevertheless, the focus on user-contributed experiential knowledge poses some practical constraints. The first one is that the form in which experiential knowledge is expressed has to be an easy and natural form to the people integrating a community of practice; otherwise, very few content will be contributed, in practice, by this people. This constraint seems to bias experience representation towards text-based content, but this again depends on the specific community of practice we are dealing with in a particular application domain. Ontology-based approaches require a highly structured representation of content, but technical communities of practice (e.g. medicine, engineering) may accept this approach if they find provided services useful.

For other users in general a text-based approach seems more suitable, but it need not be completely free text, we should be able to provide semi-structured cases where the users can textually enter their experiences. This idea leads us to the second challenging issue I'd like to discuss: the structure of experiential knowledge.

7 Forms of Experience

An important issue about experiential knowledge on the web, as mentioned before in section 4, is that *cases* as such are not already present on the web. Recalling the hotel selection example we can see there is no collection of cases of the form $(situation, outcome)$; instead we had *records of individual experiences*

[2] An overview of OWL is available at http://www.w3.org/TR/owl-features.

in the form of client reports. That is to say, we have a collection of *situations* without the *outcome*. For the task at hand, selecting a hotel, it is tempting to conceive of the *outcome* as the selected hotel: this is true for the system outputting a recommended hotel but it is not applicable to the client reports. A case in the standard sense would be a pair where a *situation* would describe the interests, preferences and constraints of a user and an *outcome* would be a hotel satisfying (most of) them. However, the client reports do not directly specify the persons interests, preferences and constraints; it is an *account* of an experience that may have been positive or negative (or something in between). Nevertheless, as I tried to show in the hotel scenario, some of this information is implicit and can be extracted: the analysis of the client records in terms of pro and con factors is a way to uncover the tacit interests and preferences of the users giving an account of their experiences.

There may be other ways to uncover the important factors in experiential accounts, since this pro and cons analysis is just an example. This leads us to the core issue in this approach: How many different *forms of experience* are there? Do we need to develop a new form or structure of experience for every new application domain? This circumstance could make impractical to apply this approach on the web at large. If not, are there a small collections of *forms of experience* that could be characterized and reused? Which are they and how to find them? I really have no answer in advance, since it is essentially an empirical matter to be settled only after trying to develop systems that reuse experiential knowledge on the web. I have some suggestions, though, as to how to proceed for developing systems that reuse experiential knowledge on the web.

The first one is trying to characterize a *form of experience* for each class of task commonly known in CBR systems: e.g. classification, regression, planning and configuration[3]. These tasks are classically differentiated by the form of the solution:

- *Classification* is a task that selects one solution from an enumerated collection of known solutions; the hotel selection scenario is thus a classification task. Variations of classification included here are: multilevel hierarchical classification and ranking of alternatives numerically or by partial ordering.
- *Regression* is a task where the numerical value of an attribute is predicted; case-base interpolation is the method of choice.
- *Planning* is a task that builds a solution composed by a sequence of actions or a partially ordered collection of actions; case-based planning has been extensively researched to deal with this kind of tasks.
- *Configuration* is a task that builds a solution composed by a network of interconnected solution elements; case-based configuration and design systems have developed techniques for this kind of task.

It seems reasonable to assume that the differences on the solution structures of these tasks imply that the corresponding experiential knowledge would also be

[3] This list does not intend to be closed or exhaustive, other tasks like scheduling etc., could be included and should be taken in to account in the long run.

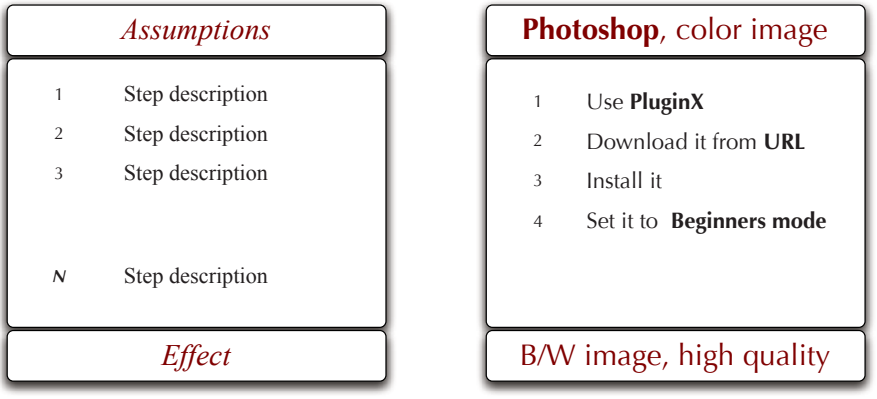

Fig. 1. Semi-structured form of experience for *How-To* tasks

structurally different. However, each class of task may have a sufficient degree of internal coherence to allow the development of experience-reuse systems applicable inside a class of tasks. For instance, the method of analyzing pros and cons in hotel client reports could be used, in principle, to other application domains whose task is a form of classification: e.g. selecting a digital camera, or selecting a B/W plugin for Photoshop. Moreover, other different techniques to reuse experiential knowledge for classification tasks could be developed. Again, only empirical evidence will determine whether the hypothesis suggested here is correct or not.

As a further example, let us consider *planning* in the context of experiential knowledge on the web. Since a plan is just a way to achieve some effect or goal performing a series of steps, it is easy to see that they are pervasive on the web, although they are not called "plans": sometimes they are called *How-To*s, but most times they are just descriptions of how to do something in few steps. Forums are websites where a large number of *How-To*s can be found. For instance, forums store numerous records of "question and answer" pairs that may be interpreted as problems and their solution-plan. A specific forum like one devoted to digital photography has both a community of practice and a shared vocabulary of terms (e.g. B/W image), verbs (actions) and proper nouns (e.g. "Photoshop"). A typical scenario is when a user asks how to perform some effect on an image and the answer is a plan of the form "*assuming you have Photoshop, you should download this PluginX from this URL, install it and then set it up in the beginner mode, you'll already have a good quality B/W image.*" Forums organize this content in a structure based on questions and answers, and thus we are expected to use Search & Browse to find and reuse this experience. Capturing this experiential knowledge from free text using NLP techniques is certainly an option, but a computationally costly one.

Another option is to design some semi-structured representation for this form of experience that, if stored on a website (substituting the questions and

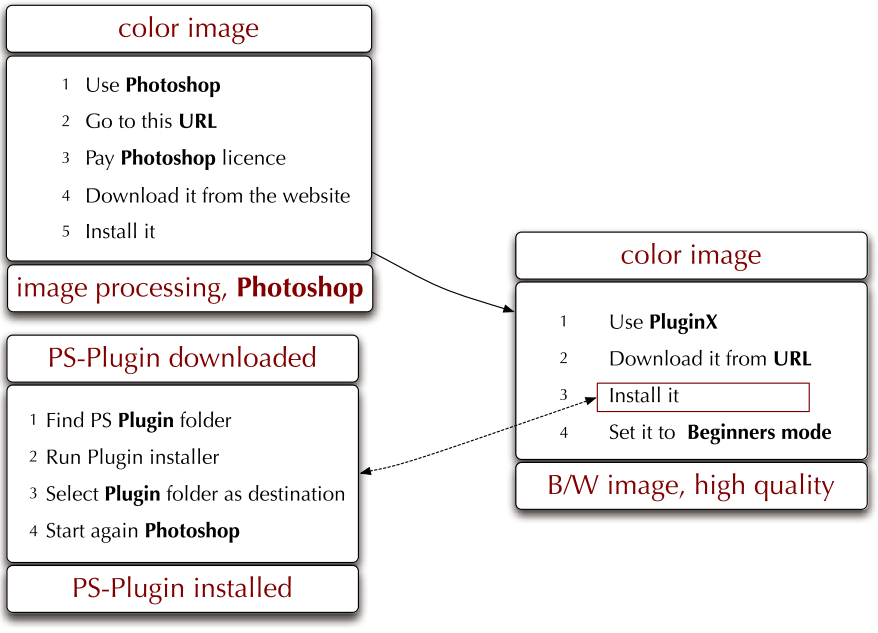

Fig. 2. Reusing experiential knowledge by combining *How-To*s

answers structure), would facilitate the analysis, retrieval and reuse of *How-To* knowledge. As a further elaboration of this scenario, consider a possible semi-structured template for *How-To* experiential knowledge as that show in Figure 1. The semi-structured template clearly separates plan preconditions (*Assumptions*), plan goals (*Effects*) and each one of the *Steps* or actions of the plan. Albeit text processing is still necessary, the previous example on *PluginX* shown at the right hand side of Fig. 1 is now more easily analyzed for the purposes of its reuse. Recall that the final user will be able to understand and perform this *How-To*, we need only enough structure to (1) allow a user to express the problem she wants to solve, e.g. "I have Photoshop and I want to transform a color image into B/W image a high quality," and (2) *recognize* that the *How-To* in Fig. 1 is a way to solve that problem.

Moreover, accessing a large repository of *How-To*s would also enable forms of *case-based plan adaptation*. Consider the situation where the user has the same goal but she does not have Photoshop. Figure 2 shows how a new plan can be generated by concatenating two *How-To*s: the first plan is one for acquiring Photoshop, while then second plan is that of Fig. 1 that uses a Plugin to achieve B/W image. Since the effect of the first *How-To* is having Photoshop, now the second plan can be safely used since the Photoshop assumption is now satisfied. Another form of adaptation is expanding a step, that is in fact a sup-plan, into its component sub-steps. Fig. 2 shows that Step 3 "Install Plugin" is not an atomic action, but can expanded into 4 steps because there is a *How-To* in the repository whose goal is to install Photoshop plugins. This form of plan adaptation should

be feasible whenever we have a large repository of plan-like *How-To*s, and it is in fact very akin to the currently fashionable idea of "mash-ups"[4] on the web.

Planning by reusing, adapting and combining user-contributed plans can be applied to a large number domains, from *How-To*s and methods to itineraries and route sheets, as long as a large repository of "action sequences" is available. The fact that these plans have been already tried by someone and were successful gives us a further advantage. The *ensemble effect* can be used on a large repository: when several methods or plans are found to achieve the same result then aggregation techniques like voting can be used to determine the one that is considered more reliable (at least inside a community of practice).

Therefore, the hypothesis put forward in this section is that several forms of experience could be defined with sufficient internal coherence so that is possible and practical to build systems for reusing experiential knowledge. The next section discusses the overall organization of such systems.

8 The EDIR Cycle

These ideas can be integrated into a process model called the EDIR cycle, shown in Fig. 3; the EDIR cycle consists of four processes: *Express*, *Discover*, *Interpret*, and *Reuse*. They should be understood as interrelated processes, not as sequential or causally dependent steps: the state of the reuse process may require changes of bias or revisions of state in the interpret or discover processes as well as the other way around.

Express. This process addresses the different ways in which experience can be expressed by a contributing user inside a community of practice. Free, semi-structured and ontology-based templates for specific forms of experience and application domains need to be developed and tested; the research goal is finding a trade-off that (a) allows sufficient structuring of the expressed experiences for automated analysis and (b) feels as a natural and unobtrusive way to express experiences for the users in a community of practice.

Discovery. This process addresses the different ways in which specific experiential content is recognised and retrieved as possibly relevant to a given query posed by a system user. The research goal is determining how to extend CBR retrieval techniques to work on experiential content integrating semantic web and/or bottom-up semantic analysis. The conditions under which the Discovery process has to work requires a fast and possibly shallow analysis of large quantities of experiential reports; the expected output is a moderately-sized collection of experiences that are (likely) relevant to the current query.

Interpret. This process addresses the different ways to build semantic interpretations of the discovered experiences. The semantics are only assumed to hold inside a community of practice. These interpretations can be understood

[4] A *mash-up* refers to a web application that combines data from more than one source into a new integrated service.

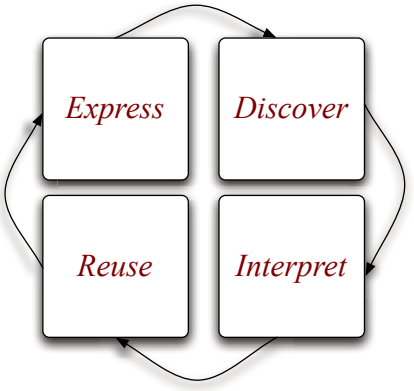

Fig. 3. The EDIR cycle for systems reusing experiential knowledge on the web

as a more in-depth analysis of the experiences selected by the Discovery process using the semantic model of the community of practice and the available domain knowledge. Several transformations are envisioned in the Interpret process: (a) eliminating a subset of discovered experiences as non-relevant; (b) transforming discovered experiences into a new canonical representation; (c) translating discovered experiences into a canonical vocabulary coherent with the one used to build the final users queries. One or several of these transformations will be used in a particular system, but the final outcome is a collection of canonical experience descriptions to be used by the Reuse process.

Reuse. This process addresses the different ways in which the experiential content provided by the Interpret process is used to achieve the goals of a user as described in a particular query. Reuse techniques from CBR may need to be revised or extended in order to be applicable in this context (e.g. case-based adaptation) but also new methods that rely on the nature of large repositories of human experience should be developed (e.g. methods based on the *ensemble effect*). Moreover there may be different modalities of experience reuse: from automated experience reuse (yielding to the user the complete solution provided by reusing experiential knowledge) to the opposite extreme where the user receives directly a small selection of relevant and reliable experiences. Intermediate modalities may perform part of the reuse process automatically while supporting the user in reuse finalization.

The EDIR cycle is a process model, so the relationship of the four processes is not sequential in an implementation of the model. Clearly, during an interaction with the final user to elucidate the requirements of her enquiry several discovery and interpretation processes may be launched and their results used to help the user narrow her options or change her preferences.

Finally, let's compare this EDIR approach with the current Search & Browse approach. The main difference is that the EDIR approach requires a *query*: a description of the kind of result needed by the system —a definition of the

problem to be solved. Only with a query it is possible to *reuse* experiences, since the Reuse process employs methods that try to satisfy the requirements of the current query using a collection of selected experiences. A second but important difference concerns the form and organization of content. The Search & Browse approach assumes the existence of just hyperlinked documents: even when some structure is present (e.g. question-answer structures in forums), this structure is not exploited to improve the results. The EDIR approach intends to characterize a particular kind of content, experiential knowledge, and it is thus concerned on how to adequately express, represent, organize, analyze, and retrieve this content.

9 Discussion

This paper is about current and future challenges on reasoning from experience. As such, I've dispensed with some formalities of the typical structure and content of scientific papers. There is not proper state of the art section, albeit sections 2 and 3 deal with the main issues on the state of the art for the purposes of this paper. There is no state of the art on natural language processing and text mining applied to the web, but this is because they are orthogonal to the purposes of this paper: they can be applied, and they mostly are applied inside the S&B approach; but they could also be used in an EDIR approach to experiential knowledge reuse.

The purpose of the paper is not presenting a specific contribution but a series of ideas that open a discussion on how to apply AI techniques, in general, and CBR techniques in particular, to the ever-growing World-Wide Web. The main idea to be opened to debate is whether there is, or is useful to conceive of, experiential knowledge on the web. I've not given a formal definition of *experience*, but my use of the term is close to the common sense meaning, and the examples presented, should be enough for a Wittgenstein-like grasp of its meaning. I found worthy of attention that trying to apply CBR ideas like reuse of past experience to the web, I've had had to abandon a straightforward notion of *case*. CBR cannot be directly applied to the web, since there are no ready-made cases preexisting on the web. However, if we understand CBR as ways of reusing past experience, we can generalize these core ideas in CBR and investigate how could we possibly reuse the experiences that people are already providing on the web.

The EDIR cycle is simply a way to organize the different issues and challenges to be addressed in developing systems for reusing experiential knowledge on the web. As such, is a tool for helping to start thinking and debating about how to build systems that reuse experience, and should be left aside when enough progress is made that shows how to proceed. I cannot claim that I can show some example system that follows the EDIR cycle, and nevertheless I can point you to the Poolcasting system, developed by Claudio Baccigalupo under my supervision as part of his Ph. D. Indeed, the Poolcasting system does not follow the EDIR cycle, since it was being developed in parallel with this proposal, and yet it shows an example of how extending some core CBR ideas we can develop a system that reuse experiential knowledge from a web community of practice.

Poolcasting generates a stream of songs that is customized for a group of listeners [10]. We needed to perform data mining processes over web communities of practice to acquire the semantics of the vocabulary of terms the systems uses. Several web-enabled information resources on the web needed to be accessed and integrated with Poolcasting to acquire a domain model. The experiential knowledge we used did not have the form of *cases*, but it is nevertheless a form of content that expresses the listening experience of the users as recorded by the music player devices. Because of this, Poolcasting is able to build, from the listening experience of a user, a model of user's musical interests that is exploited by a Reuse process.

Thus, while I cannot claim that Poolcasting is a result of the EDIR approach, it stems from the same core ideas, and as such worthy of being considered a proof of concept. The bottom-line is that I think experience reuse can be brought to the web, and the core ideas of CBR may be very useful in this endeavor.

Acknowledgements. This paper has benefited from long, engaging discussions of the author with many people, specially Agnar Aamodt, Josep-Lluís Arcos, Paolo Avesani, Klaus-Dieter Althoff, Ralph Bergmann, Susan Craw, and Nirmalie Wiratunga; all errors or misconceptions, however, are the author's responsibility. *This research has been partially supported by the Project MID-CBR (TIN2006-15140-C03-01).*

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American Magazine (2001)
2. Shadbolt, N., Wendy Hall, T.B.L.: The semantic web revisited. IEEE Intelligent Systems 21(3), 96–101 (2006)
3. Wittgenstein, L.: Investigacions filosòfiques (Philosophische Bemerkungen). Ed. Laia, Barcelona (1983) (1953)
4. Davies, J.: Semantic Web Technologies: Trends and Research in Ontology-based Systems. Wiley, Chichester (2006)
5. Benkler, Y.: The Wealth of Networks. How Social Production Transforms Markets and Freedom. Yale University Press (2006)
6. Aamodt, A., Plaza, E.: Case-based Reasoning: Foundational issues, methodological variations, and system approaches. Artificial Intelligence Communications 7(1), 39–59 (1994), `http://www.iiia.csic.es/People/enric/AICom_ToC.html`
7. Perrone, M.P., Cooper, L.N.: When networks disagree: Ensemble methods for hydrid neural networks. In: Artificial Neural Networks for Speech and Vision. Chapman-Hall, Boca Raton (1993)
8. Sunstein, C.R.: Group judgments: Deliberation, statistical means, and information markets. New York University Law Review 80, 962–1049 (2005)
9. Weber, R.O., Ashley, K.D., Brninghaus, S.: Textual case-based reasoning. The Knowledge Engineering Review 20, 255–260 (2005)
10. Baccigalupo, C., Plaza, E.: A case-based song scheduler for group customised radio. In: Weber, R.O., Richter, M.M. (eds.) ICCBR 2007. LNCS (LNAI), vol. 4626, pp. 433–448. Springer, Heidelberg (2007)