An Experimental Collaboration Tool for the Humanities: the *Procspace* System

Bernhard RIEDER[1]
Laboratoire Paragraphe

Abstract: This article describes a hybrid tool for collaborative information management and retrieval – *procspace* – that draws on different approaches in the field in order to show in what way those methods can be combined into a single application targeted for use by researchers in the humanities. Integrating data-mining models (such as the vector-space model) into collaborative text filtering can combine the semantic sensitivity of human beings with the computers aptness at processing great amounts of data. Deployment of such a tool into concrete work setting however necessitates close attention to the human and social dimensions that come heavily into play.

Keywords: Information management, hybrid approaches, collaboration, data-mining, humanities

Introduction

Despite – or rather because of – its overwhelming success and the progressive consolidation of its tinkered technical architecture, the Web is still in many ways inadequate as the global information resource that it started out to be. The use of metadata has yet to become a common practice and the *Semantic Web* has been "around the corner" for some years now. We still have to work with basically a very complex and increasingly dynamic hypertext of unstructured information stored in what could be called the biggest "content silo" in human history. As information hunters and gathers we collect documents and hyperlinks of every sort thereby all too often extending the global disorder to our own hard-drives. A lot of work has been done to advance search techniques and to design tools that create structures and hierarchies of different kinds in order to help us with orientation and understanding in the vast information landscape we live and work in. These efforts have produced interesting results but there are still large areas to explore, especially when it comes to creating hybrid applications that combine different strategies. Tapping into the collective intelligence emerging from collaboration between individuals is surely one of the most promising vectors of research.

As retrieval and structuring techniques evolve, it has also become increasingly clear that different users and user groups have different information needs and certain technical

[1] Bernhard Rieder: Ph.D. Candidate and Junior Teacher, Département Hypermédias, Université Paris 8 - UFR-6, 2, rue de la liberté, 93526 Saint-Denis Cedex 0, France; Email: bernhard.rieder@univ-paris8.fr

strategies apply better to certain patterns of work. This article therefore describes a tool, *procspace*, which has been explicitly created for supporting collaborative work groups in academia and more precisely in the heterogeneous field of the humanities, the area of the author's scientific training. We will fist establish some initial characteristics of work in this area, give then a short account of the most common strategies for information retrieval and management (IR/IM) before we present our own hybrid approach to the problem. Before concluding, we briefly discuss the design method we chose in order to show how much the success of a collaborative tool depends on the social parameters it encounters.

1. The Humanities

In the wide field of academic inquiry, the humanities[2] are commonly opposed to the natural sciences, not only because of the different objects they pertain to, but also because of their very different modes of apprehension and methods of inquiry. The area of study covered by the various disciplines inside is indeed very different from the constant objects of the natural or "exact" sciences: the humanities examine the changing forms of human existence with their inconsistencies and contradictions. Although a field can be dominated by a certain paradigm for a long period of time, this is a very unusual situation; ordinarily, ever discipline is the site of a continuous struggle between different methods and theories and the speed of change can be rapid. With paradigm-shifts come changes in vocabulary and concept space, and we should not forget that most of the humanities are highly localized, often intrinsically intertwined with national culture; English does not (yet) play the role of a unifying lingua franca like it does in the natural sciences. These are at least some of the reasons for why structured collaboration between researchers in the humanities is difficult. The ideal of the *homo academicus* [1] is still the lonely intellectual locked away in his study, and not the research group so commonplace in the natural sciences.

When trying to design tools for sustained mediated collaboration in IR/IM for this field, one has to take these characteristics into account. The humanities, seen as a heterogeneous biosphere of *communities of practice* [2], do of course feature intensive exchange between their members, but modes of collaboration are centered on informal discussion and not on structured cooperative work. At the same time, the Internet has multiplied sources of information and access to the research of others. *Information overload* specifically concerns academia and the humanities are, because of their shifting and polysemic nature as well as growing tendencies towards interdisciplinary work, even more concerned than other domains. So there is a strong need for IT based (collaborative) IR/IM but the specific requirements of researches in the humanities have not yet been catered to a lot. Resistance to computerization is considerable and tight budgets and high workloads render experimentation difficult and unattractive.

Taking into account that "a community of practice is an intrinsic condition for the existence of knowledge" [2], we wanted to design a tool that helps researchers in mastering

[2] In this paper, we use a large definition of the term "humanities": the academic endeavor of studying certain aspects of the human condition using qualitative approaches. Besides philosophy, literature, history, and cultural studies, we also consider sociology, anthropology, and political science to be part of the field.

the abundance of information on the Internet through collaboration. The basic idea was to connect IR/IM with the collective intelligence [3] emerging from sustained cooperation between individuals, while always keeping the specific situation in the humanities in mind.

2. Existing Strategies

Although IR and IM are very different from an information science (IS) as well as an engineering point of view, when looking at these techniques from a work-oriented standpoint, we see that actual practices in research integrate them into a tightly knit cycle. Researchers need not only to be able to locate interesting articles or presentations concerning their work, but they have to establish and index what resources have proven to be interesting, and which ones did not; they have to store information for quick access; this holds for both individual and group settings. We want to give a very short overview of the most common technical aides for these tasks.

2.1. Information Search and Retrieval

Seen from IS there is a fundamental difference between browsing and searching, but today all of the major search engines (e.g. Google) propose directories, and the classical directories (e.g. Yahoo) are using search engine technology. These information portals are the preferred starting point when looking for (scientific) information on the Web. The major problem with search engines and directories is that they open only a small window on the semantic content found in a scientific article; even specialized portals do not enter very deeply into actual *meaning*. Keywords and approximate categorization do not allow for high recall especially in the humanities where scientific jargon is less common than in the exact sciences and there is little consensus on the inner structure and even the main research questions in the discipline. Indexing around ten per cent of the Web, the main general search engines are still the most complete pathway into the information desert.

Data mining methods built upon semantic models, combined with good indexes (like Google's Web index) could make the Web a lot more accessible for the scientific public, but industrial strength applications (e.g. Autonomy or DolphinSearch) come at a prohibitive cost and the publicly accessible indexes of the mayor search engines are not (yet) exploitable by more complex data-mining algorithms. Only in small niches has data-mining become a means to search and/or manage information in the humanities.

2.2. Information Management

Mastering the information abundance on the Internet may involve using tools that help with organizing the data found trough the above techniques. Classic information or document management tools (e.g. *TheBrain* or ordinary database applications) or outliners (e.g. *ThinkTank* or *MORE*) help in structuring information in way that allows for precise access to already found information. The personal information repository becomes thus a structured view on the larger environment of the Internet. This rich field has been promoted

by the knowledge management (KM) community but it has found its way into the humanities on a personal rather than on a collective level.

2.3. IR/IM and Collaboration

Groupware has been a mayor direction of research since *Lotus Notes* and collective document management is now a part of workflow in the enterprise. In the notoriously individualistic academic field, computer tools for supporting collaboration have not had the same success as in the professional field. Groupware is often expensive, difficult to use and adapted to a commercial context with strong hierarchies that impose workflow and vectors of accountability on the community it sustains. But academia is not structured like an enterprise and the organization and structure of workflow is different in many ways.

New developments in the field of information management and retrieval today mostly happen at the intersection of different approaches: *Eurekster* for example combines the search engine with the power of collaborative recommendations and the authors in [5] have shown that collaboration greatly enhances data-mining performance when applied to standard text collections. Web 2.0 applications like wikis, social software and collaborative tagging are currently expanding our idea of cooperation by very simple technical means and there is no doubt that these very successful applications will leave their mark on more academic approaches.

Our own approach proposes a specific combination of the methods presented here in order to create a tool that helps a team of researchers in finding and managing scientific publications on the web. Our platform tackles the three main aspects of IM/KM: creating and discovering, sharing and learning, organizing and managing.

3. Concept and Motivation

The tool presented here – *procspace* – can be called a collaborative outliner with data-mining enhancements. The basic idea is to take the very familiar and thus less intimidating concept of “outline” (a hierarchical tree of nodes) for collecting scientific articles found on the Web, to adapt it for collaboration and to make use of the semantic structures created by users to relate existing information and to find new information on the Web. The outline provides the means to create a basic structure without imposing a specific hierarchy. Domain knowledge and structure should come entirely from the user side. The tool should enhance the personal retrieval and management performance of a single researcher as well as exchange and synergy inside of a workgroup. By relying entirely on Web technologies, there is no disruption between the search space (the Web) and the management space (the tool). There are four mayor areas we were most interested in.

3.1. Information Gathering

Procspace allows users to store resources taken from the Internet or a local computer, and to organize these nodes inside of an editable folder structure (outline). Supported documents types are the formats commonly used in the humanities: plain text, word, PDF, PowerPoint and HTML. The documents can be accessed through the server's own database or through a hyperlink pointing to their initial location. Information can also be entered through directly typed input. There is no obligatory separation between formats: a folder may contain any type of document as well as subfolders. The outline structure is flexible and can be easily changed at any moment; any node can be annotated.

3.2. Collaboration

Collaboration comes into play through the possibility for several people (e.g. a team of researchers) to work on the same outline, each person adding documents, contributing to a collective information repository. Different members of a group enhance the system with different references, thus *recommending* articles to their colleagues. Evaluation of an article's quality (through a simple vote on a five stage scale) enables users to browse different layers of collectively perceived quality. Every document can be discussed separately in a simple forum attached to the node.

These features in unison make for a form of *peer review* where part of the social organization of the process in part to the functional structure of the system. The often informal process of collaborative filtering is transposed into a sustained structure, accessible from any computer equipped with an Internet connection. The ephemeral nature of collaboration in the humanities becomes more structured or "material".

Collaborative writing is made possible through the *wiki* that can also be attached to every, or only chosen nodes in the hierarchy. People have used this feature for example to compile link-lists or to work collectively on protocols.

3.3. Data-mining Enhancements

All the activity inside of the system is based on text. First there are the primary nodes that appear inside of the outline: articles on the Web or text directly entered through the forms. Second there are wikis, annotations, and discussion threads that sticks to a primary node. The *semantic activity* of adding, discussing and writing can be seen as a pool of partly structured data that can be exploited to enhance relations inside of the system as well as to search for similar documents on the web, providing occasion for serendipity [4].

The concept proposed in *procspace* is that the semantic activity exercised in the collaboration of human intelligence is a very good starting point for machine intelligence (e.g. data-mining); we strive to create a hybrid intelligence that makes use of the unparalleled power of human beings to create meaning with the capabilities of the computer to process vast amounts of data at a very quick rate. Using an enhanced vector-space model [6] we developed a series of agents that analyze the database in order to provide a series of services: a clustered map offering an alternative view on the information

space; for every node, folder and user links to similar documents in the system as well as on the Internet; a web search feature that categorizes Google results into the systems folder structure.

3.4. Structural Openness

One of the main goals in designing *procspace* was to take into account the specific characteristics of the humanities: we wanted to avoid forcing the heterogeneous nature of knowledge in this field into a corset of pre-established conceptions about how the information space should be organized. The outline structure is therefore highly flexible and does not impose a preconceived separation of document management, discussion and collaborative annotation/writing. A node can be a document, a forum, a wiki, or all of these together. From the start, we wanted to create a semi-structured tool that would be very open in principle but gain shape during actual use in a concrete and specific work setting.

4. Architecture

Procspace is built upon a mySQL database that contains all documents in original (including HTML tags) and "clean" (stripped of all syntactic information) form. A piece of middleware written in PHP handles the flow from database to front-end and back, user management, and communication with the various agents in the system.

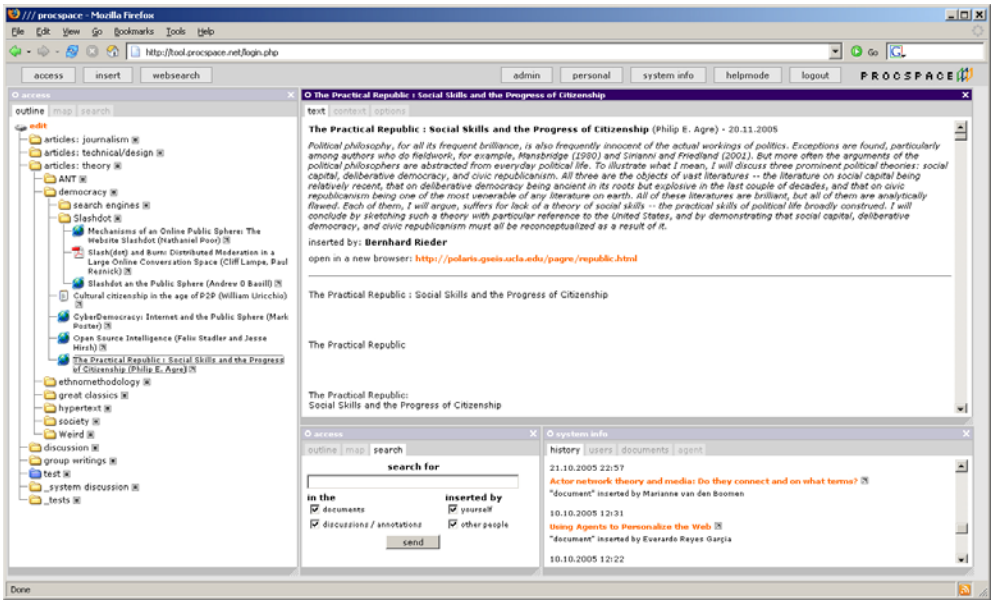

Figure 1: A screenshot of the *procspace* interface

Software agents retrieve documents from user-specified URLs, monitor changes and strip syntactic information. They also do all the data-mining work: using an enhanced vector-space model, they 1) calculate semantic similarity between documents inside of the system, thus creating a hyperlink structure parallel to the hierarchical outline; 2) draw maps based on semantic distance that propose a different kind of access to the systems resources; 3) extract keywords from every article in the system and pass the on to Google again trough the SOAP protocol; the retrieved articles are then again processed using vector-space technique to re-rank the results. This allows using Goggle's vast index without relying entirely on their heavily discussed [7] ranking algorithm.

Awareness of updates and changes in the system is generated trough a simple RSS stream that users can integrate into their favorite email-application our browser.

The interface (written in object oriented JavaScript/AJAX so that no browser plug-in is required) is based on the classic WIMP (windows, icons, menus, and pointing device) metaphor and allows every user to create his or her unique interface. This is another element in our effort not to impose too much on users, when it comes to semantic structure and workflow.

5. Design Method and Evaluation

The design method used in developing *procspace* is a combination of *rapid-prototyping* and *participatory design*. After doing a first requirement analysis and establishing initial design goals, we developed a functioning prototype that was simultaneously inserted into four (ongoing) work settings: a group of Ph.D. students in different countries, a French-German research institution, and two master programs at Paris 8 University. Feedback and observations from the different groups was (and is) constantly evaluated and when possible programmed directly into the prototype, revising requirements and structure in the process. The tool thus progressively gains shape, features and usability without loosing its open structure. Constant evaluation is thus an important part of the design process.

Although the testing phase is not yet closed, there are several conclusions we can take from our deployment experiences so far:

- A tool constructed around openness allows for very different information and collaboration spaces, but the need for preliminary and ongoing discussion between users and between users and designers is very strong.
- The higher the domain knowledge of the group, the easier it is for the users to come up with an evolving structure. For groups that are less familiar with their research field, the tools openness is difficult to manage.
- Adding data-mining algorithms to the IM tool has proved to be a valuable asset to the application without complicating the interface too much. But at the same time, such advanced algorithms insert an element of doubt in the not so tech-savvy communities in the humanities.

- While informal tools like email have become generally accepted, sustained information spaces on the Web are still strange to many of the people we worked with. This has to be taken into account when planning deployment.
- The more interesting the content gets the higher user interest and participation. The benefits of collaborative IM only become apparent when an information space has reached a certain size and quality.
- Sustained collaboration allows for new forms of working in groups in the humanities. Initial experiences are encouraging once the initial difficulties are overcome, user feedback is highly positive.

6. Conclusion

What we learn from this experiment is that besides the technical question of algorithms and engineering questions we are faced with the problem of how to design applications that make interesting use of the methods and models elaborated in IS and other disciplines over the last thirty years. Human activity on one level (researchers collecting, annotating and discussing scientific documents) can be a very promising starting point for the work of data-mining methods that enhance the performance of the human agents. Scientific work in the humanities is especially prone to be enhanced by open IM and IR techniques because the diversity of the field renders top-down classificatory approaches very difficult and imposing. Researchers in the humanities are especially sensitive to issues of power and semantic control.

The collaborative approach to IR/IM taken in our *procspace* project nonetheless shows that original combinations of existing techniques can make an interesting addition to existing work practices in the humanities.

Further research has to be conducted most importantly on how structural openness and ease of use can be brought closer together. On the engineering side, mature techniques are available, but the interfacing between tools and established work practices is still surrounded by a huge number of questions.

References

[1] P. Bourdieu, *Homo Academicus*, Ed. de Minuit, Paris, 1984.

[2] J. Lave, E. Wenger, *Situated Learning*, Cambridge University Press, Cambridge, New York, 1991

[3] P. Lévy, *L'intelligence collective, pour une anthropologie du cyberspace*, La Découverte, Paris, 1994

[4] Ertzscheid, O. and Gallezot, G. 2004. “Formalising the Concept of Serendipity in Web Searching.” Search Engine Meeting 2004. The Hague.

[5] I. M. Soboroff, Ch. K. Nicholas, *Combining Content and Collaboration in Text Filtering.* Proceedings of the IJCAI'99 Workshop on Machine Learning for Information Filtering, Stockholm, 1999

[6] G. Salton, A. Wong, C.S. Wang, *A vector space model for automatic indexing.* Communications of the ACM, 18, (1975), 613-620.

[7] S. L. Gerhart, *Do Web search engines suppress controversy?* First Monday 9/1 (2004).