

# AN EXPERIMENTAL APPROACH TO COMBINING COLLABORATION AND DATA-MINING IN INFORMATION MANAGEMENT

Bernhard Rieder  
Département Hypermédias  
Université Paris 8 - UFR-6  
2, rue de la liberté  
93526 Saint-Denis Cedex 0  
{bernhard.rieder}[at]univ-paris8.fr

## Keywords

Information management, hybrid approaches, collaboration, data-mining

## Abstract

This abstract outlines an article that describes a hybrid tool for 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. An application of this idea is shown for the field of scientific work.

## 1. Situation

Despite (or because of) its overwhelming success and the progressive consolidation of its tinkered technical architecture, the Web is in many ways still inadequate as the global information resource that it is often presented as. The use of metadata has yet to become common practice and the Semantic Web has been around the corner for some years now. We still have to work with what is basically unstructured information stored in 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 tackle this problem through 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 remarkable results but there are still wide areas to explore, especially when it comes to creating hybrid applications that combine different methods to manage and harvest the information stored on the Web in one form or the other.

It has also become clear that different users have different *information needs* and certain technical strategies apply better to certain patterns of work. This article will focus on the field of scientific work in the humanities, where textual information is still the dominant form. In science in general

(and especially in the humanities, the field of the author's training), work centers on the scientific publication as the basic unit; retrieval and management of articles and other documents important to one's own field has become a major part of our work and the increasing availability of scholarly work on the Web makes this task ever more difficult and time-consuming. We look for publications on the web, we store hyperlinks and documents on our hard-drives, and we comment them, exchange them and build our own contributions on these dense networks of existing work. Before proposing a tool targeted on this very type of work, we will discuss the classic strategies in the field.

## 2. Existing strategies

Existing technical strategies in information retrieval and management can be split into at least four categories: search engines and directories, data-mining, gathering tools and collaborative tools. Every one of those approaches has its advantages and shortcomings and it seems to be a rough consensus that a combination of methods will produce the best results. We will discuss each of them very quickly according to their importance to scientific work.

### 2.1. The search engine and directory

Issued from classic information science, those two approaches are not commonly put together, 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. By proposing a combination of *searching* and *browsing*, these information portals are the preferred starting point when looking for information on the Web. The search for related scientific work will very often pass through here. The major problem with search engines and directories is that they are basically shallow structures that capture only a small part of the semantic content found in a scientific article; even specialized portals do not enter very deeply into actual content. 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 no more than ten per cent of the Web,

the main general search engines are still the most complete pathway into the information desert.

## 2.2. Data mining

Data mining methods, 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 both on software and hardware side 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 accessible as a means to search and/or manage information in the humanities.

## 2.3. Gathering tools

Yet another way to face the "info soup" on the Web is using tools that do not actually search for information on the Web, but help with organizing the data found through browsing or searching elsewhere. Classical information/document management tools (e.g. TheBrain of intranet platforms) or outliners (e.g. ThinkTank or MORE) help in structuring information 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 kind of technology has been promoted by the KM community and has found its way into the humanities rather on a personal than collective level.

## 2.4. Collaborative Methods

Groupware has been a mayor direction of research since Lotus Notes and collective document management is now a part of workflow in the enterprise. At the same time there has been some resistance to IM platforms – often to difficult to use and to disruptive on the level of work practices. In the notoriously individualistic university space, tools for productivity have not had the same success as in the professional field.

Other classical methods for collaboration in information retrieval have been used for shopping recommendation (e.g. Firefly or GroupLens) by means of relating ones own preferences to those of others thus finding content of possible interest. Although such tools could be interesting means of exchange and serendipity [Ertzscheid 2004] also in the scientific area, I do not know of any adaptation for this area.

## 2.5. Hybrid approaches

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 recommendation and Soboroff and Nicholas [Soboroff and Nicholas 1999] have shown that collaboration greatly

enhances data-mining performance when applied to standard text collections.

This article proposes a combination of the last three methods in order to create a tool that helps teams of researchers to find and manage scientific articles on the web. Such an information management platform mirrors the three main aspects of knowledge management on a technical level: creating and discovering, sharing and learning, organizing and managing.

## 3. Concept

The tool presented here might be called a collaborative outliner with data-mining enhancements. The idea is to take a simple outline concept for collecting scientific articles found on the Web, to adapt it for collaboration and to make use of all the semantic structure inside of the system to relate existing information and to find new information on the Web. This way we hope to tackle at the same time the problem of the content silo on the Web and inside of a user's personal information collection. We hope to enhance the personal retrieval and management performance of a single researcher as well as the exchange and synergy inside of a workgroup. By using the Web technologies for our tool, we make sure that there is no disruption between the search space (the Web) and the management space (the tool).

### 3.1. The gathering tool

On this first layer, *procspace* allows a user to collect articles on the Web by the means of an editable outline or folder structure. A tree of hierarchical concepts is the basic structure of order. As in most gathering tools, the documents can be accessed in a local database or through a hyperlink at their initial destination. Users can annotate articles and a set of agents will assure monitoring and tracking of any changes in the remote document.

It is also possible to directly insert a text through a simple web form.

### 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 references to articles, thus creating a collective information repository. Different members of a research team will enhance the system with different references, thus *recommending* new articles to their colleagues.

Shared annotation allows for the discussion of an article inside of the system and adds to the deepness of information reception. 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. Both features in unison make for a form of peer review that delegates the social organization of the process in part to the functional structure of the system.

The feature for editing inserted text effectively creates the possibility of having Wiki-like collaborative writing or creating discussion streams just through annotating a call for discourse.

### 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 web form. Second there is the annotation / discussion that sticks to a primary node.

This *semantic activity* presents a stock pool of structure that can be exploited to enhance relations inside of the system as well as to search for similar documents on the web using an enhanced vector-space model [Salton 1975].

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), thus creating 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.

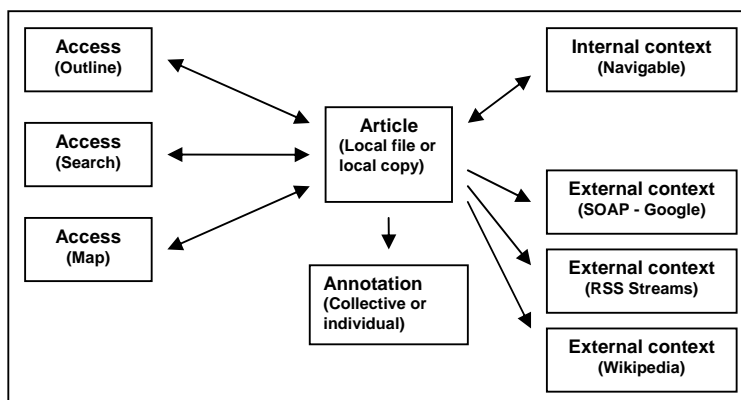
## 4. Architecture

The tool is built upon a back-end 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 through SOAP requests.

Agents (written in Java) 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 through the SOAP protocol; the retrieved articles are then again processed using vector-space technique to re-rank the results. This allows using Goggles vast index without relying entirely on their heavily discussed<sup>1</sup> ranking algorithm.

*Proospace* can also directly interface with sites equipped with RSS streams and community information platforms as Wikipedia or the Open Directory Project (DMOZ).

The interface (written in object oriented JavaScript and demoed at [http://procspace.net/bermo/files/gui\\_v6/](http://procspace.net/bermo/files/gui_v6/)) is based on the classic WIMP (windows, icons, menus, and pointing device) structure and is exemplified in this schema:



Combining human intelligence with machine approaches allow for different types of access to the collected resources inside of the system (outline, keyword search, maps based on semantic distance). It is very important that the system's users can intervene on different layers of the tools functioning. Researchers in the humanities are especially sensitive to issues of power and semantic control, as the debate about Google's ranking algorithm has shown [Gerhart 2004].

## 5. Conclusion

What we learn from this experiment is that besides the question of search algorithms we are faced with the problem of how to design applications that make interesting use of the methods and models elaborated in information science in 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 but information management and retrieval techniques because the diversity of the field render top-down classificatory approaches very difficult.

Evaluation of any kind has not yet been integrated into the project and this paper because the application is still under development and practical results and feedback have yet to be collected.

## REFERENCES

Ertzscheid, O. and Gallezot, G. 2004. "Formalising the Concept of Serendipity in Web Searching." Search Engine Meeting 2004. The Hague.

Gerhart, S. L. 2004. "Do Web search engines suppress controversy?" First Monday 1/2004

([http://www.firstmonday.org/issues/issue9\\_1/gerhart/index.html](http://www.firstmonday.org/issues/issue9_1/gerhart/index.html))

Salton, G. and Wong, A. and Wang, C.S. 1975. "A vector space model for automatic indexing." *Communications of the ACM*, 18. 613-620

Shardan, U. and Maes, P. 1995. "Social information Filtering. Algorithms for automating word of mouth." *Proceedings of CHI'95 - Human Factors in Computing Systems*, Denver. 210-217.

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

## **AUTHOR BIOGRAPHY**

BERNHARD RIEDER was born in Klagenfurt, Austria. He studied Communication, History and Philosophy at Vienna University and Information Science at Paris 8 University. He worked as a professional web developer from 1996 to 2002 and is currently teaching and preparing his Ph.D. thesis at Département Hypermédias at Paris 8 University.