Web Information Retrieval: Towards Social Information Search Assistants

Guillaume Cabanac,[1] Max Chevalier,[1,2] Claude Chrisment,[1] Christine Julien,[1] Chantal Soulé-Dupuy,[1] Pascaline Tchienehom[1]

[1] Institut de Recherche en Informatique de Toulouse (IRIT) – UMR 5505 CNRS
118 route de Narbonne F-31062 Toulouse cedex 09

[2] Laboratoire de Gestion et de Cognition (LGC) – ÉA 2043
115 route de Narbonne 31062 Toulouse cedex

{cabanac, chevalier, chrisment, julien, soule, tchienehom}@irit.fr

Searching for information is commonly an individual task which aims at solving any information need. To do that, one may go to a library, or go surfing the Web in order to find relevant information. Indeed, due to the large amount of available documents, the Web has become a favorite information source for solving daily information needs. An issue remains: the Web is in perpetual evolution; so the problem is less the existence of relevant information rather than the way users find it. One may compare searching for information on the Web with "looking for a needle in a haystack." Thus, searching the Web suffers from many limits that can be reduced by using a search assistant. Such an assistant helps the user to find relevant information on the Web. At the beginning, those assistants were principally helping each user individually. Nowadays, we are witnessing the rise of social approaches in such systems. Those latter systems help users to find relevant information by using other users' experience, shared information… Therefore, each user is helped thanks to the mass crowd.

This chapter underlines this search assistants evolution, it is organized as follows: section 1 introduces the underlying concepts and limits of traditional information search process and its application to the Web. Section 2 explains the search assistant concept by detailing their evolution from individual to social approaches. Sections 3 up to 5 present current approaches that search assistants may use to help any user to query and browse the Web as well as to improve search-related activities. To conclude, future trends for Web information assistants are discussed.

1. Searching the Web

Searching for information can be achieved through two specific modalities: querying and browsing (Agosti, 1996). Querying consists in using a specific tool such as a search engine in order to find relevant information. Browsing consists in navigating the Web thanks to hyperlinks that exist in Web pages; that is the reason why the Web is then considered as an hyperdocument. But searching for information is also correlated with many activities that must be taken into account because they are achieved during the search process. These activities are active reading, memorizing and organizing information, as well as sharing information. Moreover, it is essential to underline that a search process efficiency and success depend also on specific human factors.

1.1. Human Factors

Human factors are essential in any information searching process. Thus, as it is the case for any software, its success is conditioned by context, human factors and capabilities (Shneiderman, 1998). More precisely, in the information seeking context, two knowledge types are implied: "practical" knowledge and "domain" knowledge. So the success of any search process depends on both of them. The use of this twofold knowledge is underlined by the GVU study (1998) and by Hölscher & Strube (2000). This latter study underlines the behavioral difference between experts and newbies when seeking information according to these two types of knowledge:

- ***Practical knowledge:*** a user who wants to search for information needs to know how the Internet works and how to handle the Web itself. Thus, he has to know how to formulate and interpret a URL (Uniform Resource Locator), how to view and handle documents, how to use tools implied

in the seeking process (search engine query language…).

- ***Domain knowledge***: the most important knowledge which affects the information seeking process is domain knowledge. Indeed, it is a bit contradictory but to successfully achieve an information search process, a user must know what he is searching for or, at least, the search topic "semantic space." Thus, domain knowledge is involved at two levels of the information seeking process (Pejtersen & Fidel, 1998):
  - When the user expresses his information need by transforming his information need mental representation into a formal query which consists (sentence or keyword);
  - When the user evaluates the relevance of a retrieved document, i.e. how much the document corresponds to his information need.

So, the information searching process efficiency depends on these human factors in addition to the way users master the search process itself. To master this process, users must know how to query and browse the Web that are the two main modalities of the search process.

1.2. Information Searching Modalities

Searching for information relies on two modalities identified by Agosti (1996): querying and browsing. These two modalities are interwoven. Indeed, on the Web, a user unconsciously switches from one to another. So, when searching for information, a user browses, queries a search engine, browses again and so on…

1.2.1. Querying the Web

Querying the Web consists in using a specific tool that returns relevant information according to the queried search engine in response to a specific information need. Two approaches can be considered, namely *pull* that is mostly spread on the Web and *push* (Belkin & Croft, 1992). These approaches share main concepts but are different in the way the user interacts with systems during his search process.

a. Pull Approaches

Querying the Web thanks to a pull approach (figure 1) is very usual as it consists in using a common search engine. From the user's point of view, querying a search engine is a twofold task. First, he has to formulate his query. To do that, he must translate his mental information need into a formal query made up of words. The difficulty of this exercise is related to the way words are chosen. Indeed, words that are too general may produce results too big to be easily managed. Furthermore, words being too specific may provide no result at all, i.e. the search engine does not found any document at all. Thus, the user must have a good knowledge of what he is looking for to choose the best words to be used in a query for a search engine,. To sum up: the better domain knowledge is, the more adapted the query is, and the better the search result is. Then, the user has to manage the search results to identify the set of retrieved documents that really match his information need. This latter activity implies the user's domain knowledge that helps him to distinguish relevant information from nonrelevant one.

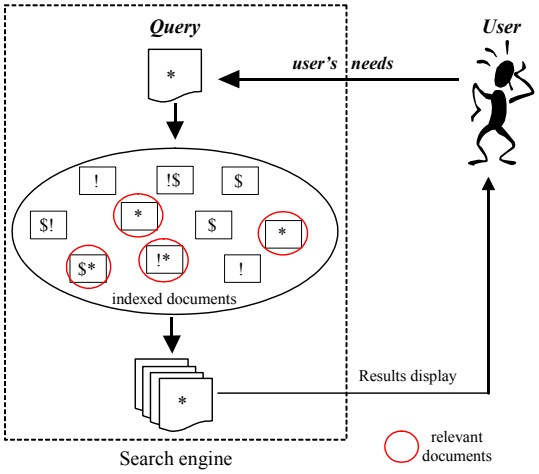

Figure 1. A search engine's common architecture

To achieve their task, search engines traditionally index documents that are generally retrieved from the Web by crawlers. The aim of crawlers is to find documents on the Web that are not yet indexed by the search engine, or that have been modified since the last indexation. To achieve that, a crawler follows hyperlinks to discover documents. Due to the hypertextual structure, search engines can only index documents that compose the so-called "visible Web" (i.e. documents that are directly accessible via an URL). The visible Web has to be opposed to the "invisible Web" which is composed of documents that are "hidden" behind a database or a form that cannot be filled in by crawlers. Bergman (2001) points out that the invisible Web is 400 times greater than the visible Web. Indexing commonly relies on specific information retrieval models and follows many specific steps (Baeza-Yates & Ribeiro-Neto, 1999). Thanks to the index from indexed documents, a search engine computes a similarity value between a user's query and any document contents. This similarity value is used to identify documents which are considered relevant according to the search engine point of view. Traditionally, relevant documents are displayed to the user through a ranked results list.

Web Information Retrieval is based on traditional pull IR techniques that have been adapted in order to improve Web search efficiency. This adaptation relies on Web characteristics such as its hypertextual structure, its size… So two pull approaches can be identified: *i*) traditional pull approaches that consider "large" volume of documents which remain "reasonable" to handle, as well as *ii*) pull approaches implemented for the Web whose volume is hard to handle. On the one hand, for traditional IR pull approaches, a collection is well-defined and generally corresponds to a homogeneous set of documents gathered on a known medium. On the other hand, for most of new IR pull approaches, e.g. on the Web, the collection is less and less easy to encompass. Documents or granules of documents are scattered, often in different formats. This characterizes heterogeneity of structure and level of abstraction. It really is the hyperlink concept which makes it possible to aggregate granules and to prevent the contents of each granule from dispersion, duplication and heterogeneity. Nowadays, the intuitive step that consists in considering that a link between two documents expresses a semantic relationship is possible thanks to hyperlinks (Hetzler et al., 1998). This assumption has sparked specific research focused on the Web, but also on Digital Libraries via the "citation" concept. The structuring of collections simplifies the problem of granules determination and the characterization of collections. In addition, the Internet and Web-related scale factor implies that, even if the user knows more or less what he is searching for, the exploited collection is only a subset of the available one (i.e. crawlers cover a small part only of the visible Web). Moreover, search results are proved to be generally larger than in any traditional approach. The hypertext concept and its underlying structure were defined for breaking sequentiality in order to approach texts on any level. First of all, it allowed the implementation of the "hyperdocument" concept, through traditional hypertexts. Then it has encouraged its application and its wide scale deployment in Web context. The deployment of this structure can be explained today by two essential motivations:

- Bringing additional information on top of a document: an author always defines a hyperlink for a particular reason. Either, he wishes to propose a page which he regards as being a reference in the field, or a page pertaining to the same site, or more generally a page containing information on the same subject, etc. However the link is generally syntactic only: determining its related semantics is a common encountered problem. Answering this question is one of the main IR concerns and also one of the Semantic Web interests (http://www.w3.org/2001/sw/);
- Capacity for sharing information and for navigating "with its own way" while following the hyperlinks. A user can be interested in a document, in a page or in a URL because of their contents or because of the documents, pages or URL they reference. This document then constitutes a starting point for navigation.

Various works have been focused on hyperlinks exploitation when searching for information, their main aim being to produce better search results. There are two kinds of approaches: the first one integrates the hyperlink concept *a priori*, i.e. at the indexing level; the second one uses links *a posteriori* to widen the result set or to better target this unit as well as to reformulate the query. Work completed in this second category generally aims to one of the following objectives:

- To reorder the documents retrieved in response to a query by taking into account the incoming/outgoing links as well as their relevance, computed according to the relevance of the linked documents;
- To use links as an alternative to the standard IR process and to propagate the activation of relevant documents towards connected documents;

- To be useful for a classification and for a categorization of a document collection.

To carry out these goals, there are two possible strategies: either to use mechanisms only based on the enumeration of incoming and/or outgoing links, or to use jointly incoming and/or outgoing links and textual contents of the connected documents, i.e. complete text or part of text surrounding the link. But initially, it should be noted that IR works that exploit hyperlinks strongly took as a starting point the studies made on scientific citations in the bibliometric field, which relates to the study of written documents connected to each other by means of citations. Various works were focused on the use of citations in order to estimate the importance or the popularity of scientific articles (Liu, 1993). The assumption underlying this type of works is that bibliographical references give credit to quoted work because they clearly influence the quoting paper. Thus, the basic idea is that citations represent judgments that authors implicitly express: if a document's author quotes another document then he thinks that it contains useful information related to the topic developed in his document. To estimate scientific articles importance, Garfield (1972) proposed a measure called "impact factor." This factor represents, for a given year, the relationship between the number of citations on the number of articles published by a newspaper, over one reference period of two years. It thus measures the average frequency with which the set of a newspaper's articles of this newspaper is quoted during a definite length of time. Citations analysis as introduced by Garfield was criticized by many authors (Hauffe, 1994), mainly owing to the fact that groups tend to quote the ones rather the others for respect related reasons rather than for relevance. Indeed, the major disadvantage of the impact factor is that it is only based on incoming citations count. In addition, it would be necessary to take into account the context of use of citations and to make a distinction between those that argue a thesis, and those that refute it. However, none of the models suggested treats this aspect.

Other measures were then proposed. They consider that citations have not the same importance, and that their influence varies according to the impact factor of the newspaper in which they appear.

However, criticisms related to Garfield's impact factor remain specific to the bibliometric field. Indeed, as Bharat & Henzinger (1998) argued, these problems occur less in Web context because the community is various and distributed, and the right of publication cannot be restricted within groups.

Citation indexing brings several advantages compared to traditional keyword-based indexing (Savoy & Picard, 2001):

- It is independent of the terms and of the language; thus making it possible to mitigate ambiguity issues related to natural language;
- Terms or sentences that describe the semantic contents of a document are prone to scientific or technical obsolescence;
- It is easier to automate because it follows a more precise syntax, e.g. URL syntax on the Web.

Regarding citation as a means of substituting or representing the contents of a document is not always objective and several motivations can invalidate the basic assumption leading to this type of practice. However, as underlined by Liu (1993), the motivations for using citations can be more complex: on top of the author's motivations, it is necessary to add the editors and referees' ones. Works on citations were largely applied to the Web. In particular, hypertext analysis has been primarily used as an alternative to the standard keyword-based IR process (Carrière & Kazman, 1997; Bharat & Henzinger, 1998; Kleinberg, 1998; Savoy & Picard, 2001). Thus, various algorithms for relevance evaluation on the Web use the underlying Web hypergraph (Agosti & Melucci, 2000). In most of these approaches, traditional IRS (Information Retrieval Systems) techniques are used to index documents (keywords) and to compute their relevance for a given query. The hyperlinks of retrieved document set are used *a posteriori* to reformulate or to rerank the result. Then we may distinguish two types of approaches, according to whether one considers only the links or that one combines the textual analysis of the documents with the analysis of the hypertext structure. The aim of Carrière & Kazman's (1997) is to reorder retrieved Web pages, it is mainly based on incoming and outgoing links enumeration. The rank of a page can be interpreted as a popularity (or quality) value based on the links surrounding this page, for example, the sum of its incoming links and its outgoing links. It is thus a measurement similar to the Garfield's (1972) impact factor. On the other hand, Brin and Page consider that links surrounding a page should not be taken into account in the same way. It is the case in particular for the Google search engine (www.google.com) which implements an algorithm that exploits the hyperlinks of the Web in a very simple way, according to the *PageRank* algorithm (Brin & Page, 1998; Savoy & Picard, 2001). As in bibliometry context, this principle consists in taking into account the fact that a page is the target of "many" hy-

perlinks (i.e. is often referred by other pages): "The more numerous hyperlinks point at a page, the more this page will constitute a good reference document." The rank of a page, called *PageRank* (Brin & Page, 1998), is computed by using a common iterative algorithm that corresponds to a principal vector of eigenvalues of the normalized matrix of the links of the Web. Savoy & Picard (2001) also proposed a method of re-evaluation of the rank of a retrieved document based on a probabilistic approach. In comparison, the algorithm proposed by Kleinberg (1998) restricts the result to the set of the retrieved pages in response to a query, that is augmented with pages pointed by (or which point at) these pages. Kleinberg defines two concepts: *Authority* page and *Hub* page. A Hub page is a page which contains links towards relevant pages; an Authority page is a page whose contents are relevant. There is thus a mutual reinforcement link between hub pages and authority pages: a good hub page is a page which points to many authority pages; a good authority page is a page pointed by good central pages. Transitively, the more a hub page points to good pages, the better it will be. In the same way, the more one authority page is pointed by good hub pages, better it will be. One of the differences between Brin and Page's algorithm and Kleinberg's one is that, for the first one, the quality (or authority) of a page passed directly from authority pages to other authority pages, without interposing a concept of hub.

Thanks to pull approaches, the user has to query information sources each time he wants to find relevant documents. A more automated way to obtain information can be seen in push approaches.

b. Push Approaches

Push approaches are relatively different from pull ones notably concerning user interaction. For pull approaches, the user has to query the system each time he wants to find documents. For push systems, the user expresses his information needs only once and the system finds automatically and regularly new incoming relevant information. Thus, in push systems, the user is informed about new relevant information each time the system founds it.

However, push systems rely on the same concepts as those developed in pull approaches (e.g. indexing, measuring query-document similarity). Indeed, they make use of the same aforementioned algorithms to index and match queries with documents. Concretely, differences between pull and push approaches (Belkin & Croft, 1992) is summed up in table 1.

Criteria	Pull approaches	Push approaches
Information need	short time	long time
Required User Interaction	high	low

Table 1. Pull approaches vs *Push approaches*

More generally, push approaches are commonly implemented through *filtering/routing* or *recommender* systems. Filtering is usually based on decision rules that select relevant documents and reject non relevant ones, see figure 2. Then, documents that are both relevant and selected constitute the recommendation set.

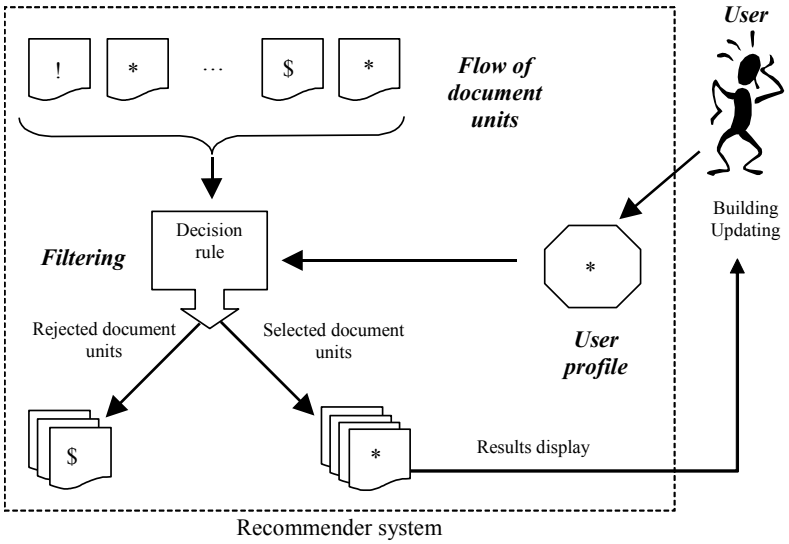

Figure 2. A common recommender system architecture

Push approaches require either the definition of a recurrent query or more generally the use of users' profiles (see section 2.1). These profiles may integrate the result of the analysis, for instance, of previous searches which can be considered as "preferences." A training stage is then necessary to build this profile which is used to personalize the search process.

1.2.2. Browsing the Web

The second modality of searching for information is ***browsing.*** This modality is based on the hypertext concept (Agosti, 1996). An hypertext consists in a network of documents linked by anchors. The hypertext on the Web is built on HTTP (HyperText Transfert Protocol) and on HTML anchors. When browsing the Web a user does not know *a priori* the structure of the hypertext he browses. To achieve successfully a search process by browsing, a user must have in addition a good practical knowledge to manipulate the Web (anchors, URLs) the best way.

Searching the Web is not so easy. It is not limited to these two modalities; it also implies many other activities that are detailed in next section.

1.3 Information Searching Related Activities

Beside previous Information Retrieval modalities many related activities are practiced during the information search process: active reading, memorizing, organizing information, spreading/sharing information. These activities are important for allowing the user to achieve a good information retrieval task.

1.3.1. Active Reading

The concept of active reading has been identified by Adler & van Doren (1972). When browsing documents or reading the snippet (e.g. keywords in context) of a retrieved document from a search engine, one has to decide whether a document is relevant or not regarding his information need. To achieve this task, a user may carry out "active reading." Indeed, active reading consists in discussing the document contents (formulating remarks, summing up the document contents ...), critical thinking and learning while reading. It can principally be supported by commentaries written down on documents. This activity directly implies the user's domain knowledge, because to be effective, active reading must be as objective as possible. Thus, the better domain knowledge is, the better the evaluation of the relevance of a document is.

1.3.2. Memorizing & Organizing Documents

Memorizing relevant information is a starting point and an ending point of the information search process. Indeed, a user can start the process from an already known document (search engine, portal ...) and can memorize new documents for a further use. This activity is not limited to the information search process. For instance, when dealing with paper documents, Kaye et al. (2006) have shown that people archive documents for many reasons, one of the most important being the need to find interesting documents later again. This field study does not reveal a universal organizational practice: everybody has his own strategy that works for him. Therefore, personal paper archives are organized in many ways. This observation can also be applied to digital documents.
These memorized documents may also be used as a starting point: a user can start browsing from a specific and already known document or from the result of a search engine.

1.3.3. Spreading/Forwarding Documents

When searching for information, one related activity consists in spreading documents found. Indeed, each time a user finds a document, he may suppose that one of his friends or colleagues may be interested in it. Thus, the user may spread or forward documents to specific users. This activity is difficult because it must not be intrusive for any user. Furthermore, this activity implies that the user who wants to spread information knows other users' information needs or up-to-date interests. Finally, this activity is highly resource consuming (in time, for instance) because spreading documents is commonly achieved manually *via* email sending. This activity is important because it can be considered as a supplement to the information searching process (Table 3).

71.0 %	Via search engines
9.8 %	Proposed by friends or colleagues
8.5 %	Quotidian books
8.4 %	A link in another Web page
8.1 %	Randomly, via browsing
3.6 %	TV advertisement
3.3 %	Books referencing Web pages

Table 3. How a user arrives on a web site? – CommerceNet/Nielsen Media – July 1997

1.4. Limits of Information Searching on the Web

Searching for information suffers from many limitations that have been identified following many IRS fails (Baeza-Yates & Ribeiro-Neto, 1999; GVU, 1998). These main limits are:

- The impossibility of finding relevant documents,
- The difficulty of querying a search engine,
- The difficulty of using the search engine query language,
- The fact that only the visible Web is really indexed,
- The difficulty of managing retrieved documents from a search engine,
- The impossibility of finding a document which has already been encountered,
- The impossibility of returning to a visited document (history),
- The impossibility of knowing where the user is located in the hypertext (disorientation),
- The impossibility of visualizing the hypertext structure,
- The cognitive overload occurring when browsing the Web caused by the hypertext structure memorization,
- The instability of Web documents (dead links),
- The difficulty of organizing memorized documents when they gets numerous.

To limit these issues, many solutions have been proposed in order to help the user to search for relevant information. Such approaches are implemented in search assistants. Next section presents such systems while underlying the rise of collective/social approaches that aim at helping users searching the Web.

2. Search Assistants: Towards Social Information Search Assistants

Searching the Web via previous traditional approaches is not fully efficient because they do not solve the user's confusion issue when facing the flooding mass of available information. Thus, to help any user to find relevant information on the Web, search assistants have to take into account:

- The users' interests through the concepts of "profile" and "context";
- The users' specific wishes concerning preferences for documents which may be exploited by a relevance feedback algorithm, for instance;
- The users' behavior. Based on search engines uses and users' practices for instance, statistical analyses may be carried out. They aim at characterizing the users and their needs as close as possible (Jansen et al., 2000). On the Web nowadays, it is possible to observe users' behaviors in a relatively simple way and in real time. For example, it is easy to take into account:
    - The most frequent queries for a search engine;
    - The average time that a reader spends on a page or a document;
    - The navigation paths followed from a page or a given document;
    - The language used by the users to formulate their queries, as well as the language of the documents that they read…

These practices are very frequent for the e-marketing domain or for direct marketing to target potential customers. They may however be of a great interest for IR itself, for example:

- Query analysis makes it possible to characterize a user or a group of users' most approached;
- Frequency of result selections enables to penalize documents that are not often–even not at all– selected, and thus allows to optimize the relevance of the other turned over results;
- Each net surfer's navigation analysis through statistics makes it possible to define "interest areas" by computing similarities between sites (competitor or not) and to suggest new destinations likely to interest users.

In fact, users' behavioral analysis may be a base for profiles definition, i.e. users' representation. So the "user profile" notion is detailed in next section. Indeed, to help the user to find relevant information on the Web, these approaches take into account the user, his activity and his information needs. Then, specific approaches aiming to help a user to improve the search process are presented meanwhile focusing on the rise of social/collective approaches.

2.1. A Required Step: User Profiling

As users' satisfaction is the main goal of search assistants, describing the users precisely, the most faithfully and with details is a key step towards that goal. Generally, a user profile is described by a set of characteristics that identifies or represents him. User profiling can be studied under various aspects: *users' profiles classification*, *profiles representation models*, *profiles definition methods*.
According to the semantics of a given characteristic, we can classify users' profiles as defined in table 2 below

Users' profiles classes	Semantics
Users' needs profiles	Profiles describing the users needs or interest centers.
Users' judgments profiles	Judgments of users on a set of documents.
Users' demographic data profiles	Profiles describing different demographic data of the users: name, gender, age, profession, address, and so on.
Multi-criteria profiles	Profiles describing various characteristics of the users: needs, judgments, demographic data, and so on.

Table 2. Classification of users' profiles

The description of a user profile generally follows a given model. We discern two widely used models for users' profiles representation: the *attribute-value model* and the *hierarchical model*.
The *attribute-value model* describes a user profile by a set of independent attributes bounded to an atomic value (string, numeric, date…). For instance, one may describe a user's demographic data with the attribute-value model as follows: *(name, Peter), (gender, male), (age, 18), (job, student)*. The first element of each previous pair represents the attribute and the second element describes the value bounded to this attribute. The attribute is considered as a key so in an attribute-value model two attributes with the same name cannot exist at the same time.
The *hierarchical model* organizes various characteristics (or attributes) of a user profile as a tree where leaves are bounded to values representing the contents of the profile. The hierarchy is a way of defining a relationship between the different characteristics. Hence, each non-leaf node represents a class of attributes. Moreover, in a hierarchical model two attributes with the same name but with different access paths in the hierarchy can exist at the same time.
More specifically, there are *contents representation models* of profiles that somehow define their exploitation framework. Generally, contents of users' profiles are represented by a *list of atomic values* or by a *list of weighted values* (Korfhage, 1997). Atomic values lead to a form of database matching when comparing different attributes while weighted values allow a more fuzzy matching that evaluates a degree of similarity between different attributes instead of a binary similarity value. Weighted values are mostly used in IR and the cosine formula is generally used to measure inter-profile or query-profile similarities (Baeza-Yates & Ribeiro-Neto, 1999)…
Manual methods are often used to define profiles; to do that a user generally fills in a form. On the other hand, automatic or semi-automatic methods may also be used to define attribute contents of profiles such as indexing, clustering, profiling and stereotypes approaches:

- *Indexing* consists in selecting the keywords that best characterize a text (document, query ...). For each

keyword a weight is calculated by using *tf.idf* like formulas (Baeza-Yates & Ribeiro-Neto, 1999). So one may define the given user's interests by indexing the set of documents that he has visited, saved or judged (Godoy & Amandi, 2006).

- *Clustering* or *machine learning* (Leouski & Croft, 1996) consists in identifying objects classes based on similarity of their characteristics. Clustering tries to minimize variance inside a given class and to maximize this variance between classes. Clustering result is then a set of heterogeneous classes with homogenous contents. Hence, one may create users' profiles by applying clustering methods on the set of document contents they saved, judged or visited in order to discover a user's interests or topics.
- *Stereotypes* (Shapira *et al.*, 1997) consist in pre-defining classes and characteristics of these classes. Documents or users are automatically bounded to a given classes according to their contents. The stereotypes approach is a kind of clustering and is mainly used for defining users' groups.
- *Profiling* (Cho *et al.*, 2002) consists in tracking the user during his different log sessions and in analyzing his behavior. Profiling helps to find documents saved or judged by a user. Therefore profiling is mostly used in electronic commerce in order to identify which kinds of products a user is looking for and then recommend him items that meet these needs. For that purpose, profiling generally analyses clicks on products, products saved in a shopping basket, purchase of products…

Note that the analysis above on users' profiles (classification of users' profiles, associated generic models, contents representation models, defining profiles methods) is the same for either individual user profiling or for users' group profiling. Furthermore, on top of the user's interests, a user profile can incorporate more behavioral information.

Thanks to this representation of users, search assistants can implement algorithms which assist the user during his search process. Two types of search assistants are considered in this paper:

- ***individual search assistants*** which help a single user by exploiting information concerning this user only,
- ***social or collective search assistants*** which help a single user or a group of users by exploiting information concerning the user in addition to information about other users (experience, shared information…). For such systems, each user is benefiting from the mass crowd.

The actual rise of social search assistants can be explained by the evolution of the Web as a whole. The new Web also called "Web 2.0" encourages such types of systems.

2.2. The New Web: the Rise of Collective Search Assistants

We nowadays attend the birth of a new Web which is meant to be more open, as an evolution of the previous "Web 1.0, ". (O'Reilly, 2005) has named it "The Web 2.0." Concretely, this new Web goes beyond its initial role of "a worldwide electronic library" as it shifts to an alive, dynamic and interactive space where people may get involved and benefit from it. Actually, Web users go from a passive reader state to an active role of contributor. In fact, this change of era is twofold: a technical facet and a social one characterize it. On the one hand, the Web 2.0 may be perceived as a replacement of the Web 1.0 techniques by services oriented techniques. On the other hand, it represents a new network made up of social interactions. The blogs, RSS (Really simple Syndication) feeds, the wikies… belong to this tendency. People also call this new Web "The Read/Write Web" to color the fact that it is no more exclusively oriented towards publication but also supports a strong collaboration between their users (http://www.readwriteweb.com/). This Web 2.0 repositions the user and the relationships he has with others at the heart of the Internet. According to Tim O'Reilly the key to success in this evolution of the Web lies in "the Collective Intelligence". Moreover the economic stake seems important since many industrialists and leaders on the market of data processing also seem to be interested in the fact that people are more efficient when they work collectively. For example, it is the case for Microsoft which has published in France a document about collective effectiveness which gathers analyses, testimonies and reflections carried out by experts of the domain (http://www.ec2006.blogs.com/). In less than 10 years, we passed from static Web pages to a Web made up of objects, people and services, bound by a central aggregator: the individual.

In this context, information searching techniques have also evolved. Indeed, the information retrieval task was first of all approached as an individual step. Each new IR task was regarded a long time as an isolated act. But today the collaborative aspects in the processes of information retrieval attract a growing interest. Indeed, if one observes the most frequent situations where users need to seek information, one realizes that

they are generally in social, organizational or professional contexts. In these contexts, users act as groups and need to share their experience, as well as to retrieve information (Karamuftuoglu, 1998; Fidel et al., 2000; Talja, 2002). Therefore, these observations guide a new research aiming at helping users in a setting of information seeking and so in real-life environment (e.g. communities of academics, scientists, engineers…).
More recently, Hansen and Järvelin (2005) demonstrated that collaborative IR is very frequent, even more frequent than expected. Thus, they give the following preliminary definition of Collaborative Information Retrieval (CIR): "*CIR is an information access activity related to a specific problem solving activity that, implicitly or explicitly, involves human beings interacting with other human(s) directly and/or through texts (e.g., notes, figures, documents) as information sources in a work task related information seeking and retrieval process either in a specific workplace setting or in a more open community or environment*." Moreover, they analyzed IR activities according to how and what collaborative activities are involved. They identified two types of collaboration in IR activities:

- ***Document-related collaborative activities***: this concerns creation, sharing or re-use of documents. These activities include:
    - Sharing and re-using documents (created or retrieved in the same community);
    - Creation, sharing and re-using contextual relationships (annotations, citations, references);
    - Sharing and re-using judgments, decisions or opinions (objective or subjective ones);
    - Sharing representations of information need (query terms, query structures…);
    - Sharing the history of information objects such as logs, links, bookmarks or documents themselves.
- ***Human-related collaborative activities***: this concerns the use of knowledge possessed by other humans (individuals or groups). These activities can be explicit or implicit, and most of the time they are in verbal form. Among those activities we can quote:
    - Sharing tasks (cooperation) and/or sharing sub-tasks (division and distribution of tasks);
    - Sharing search strategies, search terms, or classification codes…
    - Sharing or asking for external and/or internal expertise. Users of a same group (or having the same goal) may be asked for domain specific knowledge, as well as for information retrieval specificities;
    - Sharing internal experience;
    - Communicating and sharing advices, personal and subjective opinions…

In the same way, various studies undertaken in the information science domain come to corroborate these aforementioned observations (Sonnenwald & Pierce, 2000; Prepok, 2002). It is then obvious today that the information retrieval related activities may be improved by only taking into account the fact that any user may profit and benefit from any form of collaboration within the framework of given activities. Moreover, it is argued that "*the fundamental intellectual problems of IR are the production and consumption of knowledge. Knowledge production is fundamentally a collaborative task, which is deeply embedded in the practices of a community of participants constituting a domain*" (Karamuftuoglu, 1998). From now on, current advances in networked systems cross the boundaries of researches on IR: they do not only concern technical problems but also human and intellectual aspects, and then a better understanding of the social aspects of information retrieval.

In this context, it is interesting to underline different approaches that can be implemented to help the user to improve his search process via search assistants. In next sections, these approaches are detailed taking into account each type of assistants, that is to say individual and social assistants. Next sections follow the different search process steps to present different search assistants.

3. Querying the Web Assistants

3.1. Pull-based Assistants

3.1.1. Individual Pull-based Assistants

To help a user to query an IRS, many approaches may be implemented at each step of the querying process: a

search engine selection, the formulation of the information need (query), the management of search results. Thus, the provided help concerns the important steps which condition the quality of the search process.

a. Selecting the Suitable Search Engine. On the Web, many search engines can be found: some of them are domain specific, other one is generalist. When searching for information, a user has to select the right search engine to get the most relevant documents. For instance, *GlOSS* (Gravano et al., 1999) indexes several search engines. When searching for information, a user can first query *GlOSS* to find the most adapted search engine. Then, the user can directly query the selected search engine. The limit of such approach is that *GlOSS* must have access to the search engines' contents. Moreover, search engines suffer from the limited overlap between search results for the same query. That is to say, for a same query, different search engines may not retrieve the same document set. To limit this issue, meta-search engines have been implemented.

b. Selecting Multiple Search Engines. Meta-search engines aim at querying multiple search engines for the same query and to synthesize search results. *MetaSearch* (www.metasearch.com), *InternetSleuth* (www.isleuth.com) or the well-known *Copernic* (www.copernic.com) are examples of tools that use such approach. A user first selects the list of search engines he wants to query. Then, the meta-search engine queries each search engine while eventually adapting the query. Lastly, search results are synthesized and displayed in a single ranked list.

Dreilinger & Howe (1997) propose a mixed approach that features search engine selection associated with meta-search.

Even if these approaches allow the user to query the most suitable search engines, a main difficulty remains: query formulation.

c. Formulating a Query. Helping users to formulate their queries is gaining a great importance, more particularly in Web context. When the user formulates his information need by a query, the chosen terms for his query have an influence on the response of the system. However, generally, the user formulates his queries with its own vocabulary which may not strictly correspond to the one used to index the relevant documents of the queried collections. So, some approaches like *WebCluster* (Mechkour et al., 1998) offer query formulation via a mediating system. Such systems display the user many sets of documents. Then, the user selects the set of documents which corresponds to his information needs. Thanks to this set of documents the system automatically generates a query according to the indexing terms contained into these documents. Then the user may use the generated query to interrogate any search engine. Thanks to this query or thanks to his own query, a user is assisted while adapting or reformulating his query in order to find what he is really searching for.

d. Reformulating a Query. In order to select the maximum of relevant documents while limiting the noise (i.e. too many nonrelevant documents retrieved), the user should choose known relevant terms from the indexing language. This task proves to be difficult insofar as, in general, on large corpus, it is impossible to know the indexing language used. Indexation, and in particular its exhaustiveness, thus have a direct incidence on the quality of the answers of the system. One can conclude from it that, taking into account increasing volumes of the collections, to find relevant information by using only the initial query is a quasi-impossible operation. The automatic reformulation of the initial request of the user is a means of mitigating the problem and of helping the user to target his need. The query reformulation is a process aiming at generating a new query more adequate than the one initially formulated by the user. This reformulation makes it possible to coordinate the language of research (used by the user in its query) and the indexing language. Consequently, it limits the noise and silence due to a bad choice of the index terms in the expression of the query on the one hand, and the gaps of the indexing process on the other hand. We distinguish mainly two approaches to reformulate queries, (1) according to whether they use predefined term associations, or (2) the relevance and not relevance of the documents retrieved in response to an initial query. The two principal techniques used are respectively query expansion and relevance feedback.

(1) Query expansion is based on the following principle: the simple comparison between the contents of the query and the documents of the collection does not make it possible to have all the documents corresponding to a given query. So, some relevant documents remain not retrieved. Research tasks proposed to reformulate the initial query by the addition of the semantically close terms. Those terms may result from:

- Studies on the natural language (morphological alternatives ...). It is thus possible to add to the query

the morphological alternatives of the various terms employed by the user. The goal of this mechanism is to ensure the restitution of the documents indexed by alternatives of the terms composing the query. Within this framework, one uses stemming and truncating algorithms,

- Statistical studies and analyses on the contents of the documents of the base. One can thus choose to add a certain number of the most relevant terms of the selected documents, or to preserve of it only one number limited among the initial and added terms.

In the same way, another method proposes to add terms close to or terms associated with those of the query. It is a question of seeking inter-terms associations. In this direction, various research tasks were undertaken. One distinguishes the manual methods, using for example predefined thesaurus (directed indexing) and methods entirely automatic such as:

- The computation of the contextual links between terms;
- The computation of correlation matrix between terms;
- The automatic classification of terms;
- The automatic classification of documents.

These associations are generally created automatically and based on the co-occurrence of the terms in the documents. The inter-terms links reinforce the concept of relevance of the documents compared to the queries. Manually established associations generally represent relations of synonymy and hierarchy. The manually built thesauri are an effective means for the query expansion. However, their construction and the maintenance of semantic information that they contain are expensive in time and require the recourse to experts of the considered fields. Currently, the most used terminological resources, as well on the Web as on dedicated systems, are the *SynSets* of *Wordnet* (http://wordnet.princeton.edu/).

(2) Relevance Feedback takes into account the user's relevance judgments on the documents which have been retrieved in response to its initial query. The user can provide judgments of relevance with regard to the retrieved documents by stating those that it considers relevant and those that it considers nonrelevant. These judgments are then used to reformulate the query. The fundamental principle of relevance feefback is to use the initial query to start research, then to modify this one starting from the judgments of relevance and/or non-relevance of the user, either for reweighting the terms of the initial query, or to add to it (resp. to remove) other terms contained in the relevant (resp. nonrelevant) documents. The new query obtained, at each feedback iteration, makes it possible to correct the direction of research in the direction of the relevant documents. The approach described in (Rocchio, 1971) proposes to repeatedly derive the optimal query starting from operations on the relevant documents and nonrelevant documents. The idea is that the relevant documents have terms associated with the terms of the initial query; their semantic weights are then increased by the addition of the relevant document terms. Conversely, the weights of the terms contained in the nonrelevant documents are decreased. In the same way, the terms absent from the initial query are added. Then Rocchio's work was extended by Ide (1971) to develop two other strategies which appeared more powerful according to various studies (Salton & Buckley, 1990). The first strategy is the basic Rocchio's formula without standardization taking into account the numbers of relevant documents and nonrelevant documents. The second strategy is similar to the first one, but makes it possible to limit the feedback starting from the nonrelevant documents by using only the k first nonrelevant documents in the ordered list.

At this stage, the user must, first of all, be able to judge retrieved documents and thus, it is necessary that he can manage search results.

e. Managing Search Results. When querying a search engine, an important step is to manage the search results. To do this, the user has to analyze search results to identify which documents are relevant to his information need. To help the user to identify what he is searching for, a single ranked list may not be suitable. Indeed, this ranked list is not efficient when the number or retrieved documents are too numerous and when a global view of the retrieved documents is necessary. To help the user to identify relevant information, various visualizations have been proposed. These visualizations are also called Visual Information Retrieval Interfaces (VIRI). A classification of such interfaces has been proposed in (Zamir, 1998). Figure 3 shows an example of VIRI extracted from the *Easy-DoR* software (Chevalier, 2004) that represents retrieved documents in a 3D space. Each colored plot corresponds to a set of retrieved documents having the same combination of query keywords.

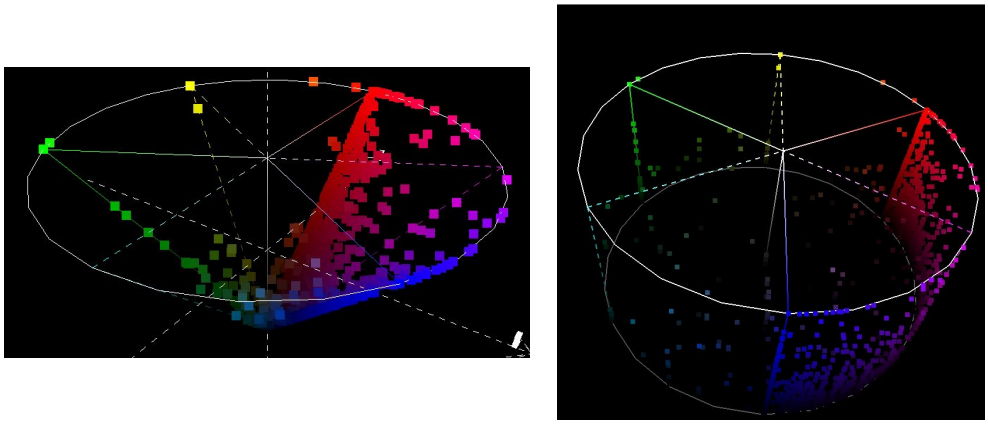

Figure 3. VIRI proposed in EasyDoR

In addition to such visualizations, clustering approaches as *Scatter/Gather* (Cutting et al., 1993) may be applied to limit the number of displayed documents and to make search results more interpretable.

3.1.2. Social Pull-based Assistants

The large majority of social assistants concern distributed IRS and peer-to-peer (P2P) networks. The guiding principles are based on the fact that it is more judicious and easier to create and share knowledge at a local level, in a definite context where the relevance of information is better controlled (Wu & Aberer, 2004; Klemm & Aberer, 2005; Jin et al., 2006; Kwok, 2006). These works all focus on decentralized IR and integrate contextual information sources.

Peer-to-Peer information retrieval like *HumanLinks* (Memmi & Nérot, 2003) is increasingly receiving attention (Yang & Garcia-Molina, 2002; Kwok, 2006). Indeed, like peer-to-peer systems, P2P information retrieval systems have emerged as an appropriate way to share and re-use huge volumes of documents. Yang and Garcia-Molina defined P2P systems as "*distributed systems in which nodes of equal roles and capabilities exchange information and services directly with each other.*" According to Milojicic et al. (2002), the main benefits of a P2P approach include: "*improving scalability by avoiding dependency on centralized points; eliminating the need for costly infrastructure by enabling direct communication among clients; and enabling resource aggregation.*"

To sum up, information seeking and retrieval can benefit from P2P systems features in order to bring some help to information seekers and to optimize their tasks. Information searching is then decentralized and realized either locally or routed (duplicated) towards other peers. Each peer must maintain an information repository available for sharing. Several studies report different experiences in P2P information retrieval. In a social context the most significant work is discussed in what follows. Kwok (2006) describes some queries and searching trends in P2P networks. Wu & Aberer (2004) and then Klemm & Aberer (2005), Jin et al. (2006) study different strategies for searching, they have implemented and experimented them on P2P networks. The main problem remains: whatever the basic strategy on each peer, the aggregation of the partial results obtained on each peer in a global result (global ranking or global relevance evaluation) is quite difficult to implement.

Wu & Aberer (2004) revisit the *PageRank* algorithm to apply its principle in a decentralized way. They propose to compute a *SiteRank* on the pages of each site and then to aggregate the rankings from multiple sites.

They argue that while decomposing the task of global page ranking computation to distributed participating servers (in a decentralized search system), their strategy can overcome the missing of a global view. Moreover, spamming becomes very difficult with *SiteRank* computation.
Later on, Klemm & Aberer (2005) study a full-text retrieval strategy on P2P technology. Each peer acts as an independent IRS with its own collection, its own indexes (local term vocabulary). Local term frequencies can be computed and then global term frequencies can be inferred.
Jin & al. (2006) propose an information retrieval strategy based on semantic small world in an unstructured P2P document sharing system. In such architecture, each peer maintains its own semantic representation. The computation of the similarity between two peers is based on a formula derived from the cosine measure. Short-range links and long-range links are identified to construct a semantic small world, as well as to reduce network traffic.
In another way, several works were interested in the semantic description of the sources in distributed systems or in P2P systems, among them Haase et al.'s studies. Haase et al. (2005) proposed to organize knowledge (as metadata description) within communities by the use of ontologies. They have extended the *Bibster* system by integrating an ontology learning process. The ontology is built on each peer from the users' local repositories. Thus, each ontology reflects the personal interests of local users. Such learned ontologies allow systems to create personalized classification and then personalized information retrieval. So in a social point of view, the common point of these decentralized strategies, and their main advantage too, is that users (and by extension, peers) may limit their search to some reliable sites. The target can be sites they already have identified, or sites which were recommended to them. In that way, the main help that can be brought to any user is to guide him in his search step, by providing local terminological resources, as well as topological maps of search spaces. It is still necessary to take into account the collaborative and social aspects to construct and to broadcast these semantic resources.
Collective pull approaches can also use semantics stored in annotations. Thus, Fraenkel and Klein (1999) argue that it is worth exploiting such human-contributed contents to improve Information Retrieval. Merging annotations corresponding to a given query is proposed in (Sannomiya et al., 2001). More recently, in the Digital Library area, Agosti and Ferro (2005) have been considering people's annotations for re-ranking results of search engines. Collective annotations are also considered valuable by Frommholz and Fuhr (2006); they even model exchanged emails as discussion threads for the TREC Enterprise Track.

From the user's point of view, the management of search results can be improved thanks to collective aspects. A relevant example is illustrated in the system *VR-Vibe* (Benford et al., 1995) which proposes a collaborative space containing all queries and search results of a group of users. Users can interact together via annotations associated with retrieved documents.

3.2. Push Approaches: *Filtering or Recommender Systems*

3.2.1. Individual Recommender Assistants

Filtering or recommending information is a passive task since the user does not explicitly express his needs through a query formulated in a more or less structured natural language (as it is the case for pull-based search engines). User's needs are stored in his profile and are then compared to available information in order to find which one best corresponds to him. For an individual access to information, filtering is generally content-based.
Individual content-based filtering uses the description of information contents in order to determine to which individual users' needs profiles they corresponds. Their recurrent interests that are described by a list of weighted keywords (Korfhage, 1997), see section 2.1, define the users' profiles. These profiles are obtained by indexing users' information saved or judged, for instance (Bottraud *et al.*, 2003).
On the other hand, one can also recommend a given user a document similar to another one, previously judged interesting by this user. In this particular case, the user profile is described by his judgments passed on several documents and the comparison or matching process is performed between two information descriptions where one has been judged by the user and the other one is the new candidate information (not yet judged) for recommendation.
Content-based recommendation may also be based on parts of documents. For instance, one may use parts of

documents that contain words of the user's information needs in order to find other relevant documents (Teevan *et al*., 2005). All the same, parts of documents that have been highlighted by a user can also be used for finding and recommending documents that correspond to him (Price *et al*., 1998).
Suitor (Maglio, 2000) is a representative system that studies the user's environment and actions to build his profile. The approach innovates by various functionalities. In particular, it uses a scrolling window making it possible to post the information suggested by *Suitor* on bottom of the screen. It proposes especially a system enabling to follow the user's glance via a camera in order to identify parts of the screen which he is looking at.

3.2.2. Social Recommender Assistants

There are several types of collaborative filtering or collaborative recommending (Montaner *et al*., 2003; Malone *et al*., 1987) such as: social filtering, demographic filtering, content-based collaborative filtering.

a. Social Filtering. Social filtering uses judgments (or feedbacks) of a set of users about a set of documents in order to perform recommendations. To do that, a similarity measure is computed between the users' judgments (Goldberg et al., 1992; Konstan et al., 1997). This measure may help to determine if a judged document corresponds to a given user. In social filtering, the description of document contents is completely ignored. Table 4 represents samples of users' judgments in a social filtering framework. The "+" symbol means that the given document is interesting for the given user, the "–" symbol means that the document is out of interests for the given user and the "?" symbol means that the document has not yet been judged and hence is a good candidate for recommendation. According to table 4, the users named $User_1$ and $User_3$ can be considered as similar since they have made the same judgments on the same documents. One may hence recommend $Document_4$ to $User_3$, since the $User_1$ has already judged it as interesting.

Users	**Judgments on documents**			
	$Document_1$	$Document_2$	$Document_3$	$Document_4$
$User_1$	+	–	+	+
$User_2$	–	–	–	?
$User_3$	+	–	+	?

Table 4. Samples of judgments profiles of users for social filtering

b. Demographic Filtering. Demographic filtering uses demographic data (gender, age, profession, address, etc.) in order to create groups of users and make recommendations within a given group (Krulwich, 1997). To do that, personal judgments are classified according to individuals' demographic data. This categorization allows determining which type of information is appreciated by a given type of user. The categorization either may be manual or based on users judgments, for instance, in order to deduce the type of users (or group of users) to which information corresponds (Pazzani, 1999).

c. Hybrid Filtering or Content-based Collaborative Filtering. The previous collaborative filtering methods are not exclusive and can then be combined. Therefore, different hybrid methods have been developed by combining different kinds of filtering methods (Good et al., 1999; Pazzani, 1999). The use of hybrid approaches improves the results' relevance of the various filtering systems introduced previously by mitigating their limits (Balabanovic & Shoham, 1997) which are: over-specialization in individual content-based filtering; users' judgments on a document which is a very time consuming task…

Hybrid filtering methods may also be called content-based collaborative filtering, since they generally combine content-based filtering with social or demographic filtering as follows:

- users' similarities can be deduced thanks to their profiles needs constructed with the information contents that they have already judged (Pazzani, 1999). So, in order to identify groups of users, one may no longer consider similarity between users' judgments only. The point of this type of hybrid filtering is that it allows making recommendations to a new user by determining his group from his profile needs. In pure collaborative filtering, one should have to wait until this new user had made enough judgments on information in order to associate him to a given group and then recommend him some documents. This is called the "cold start problem" of pure collaborative filtering systems;
- all the same, documents similarities can be deduced by comparing their respective contents instead of

comparing judgments made on these documents (Yager, 2002). Hence, one can recommend information if its description is similar to another one that has been already validated (judged interesting) by the user. An interest of this approach is that it is possible to recommend information that has not yet been judged. Whereas in pure collaborative filtering, an information has to be judged at least once to be recommended.
In point of fact, collaborative filtering has as a principle of exploiting the evaluations made by a group of users upon certain documents in order to recommend these same documents to another group of users. All the users of a collaborative filtering system can benefit from the evaluations of the others by receiving recommendations for which the closest users gave favorable (positive) judgments. That does not require that the system has a process of extraction of the document contents. Moreover, the user is able to discover various interesting fields, because the principle of collaborative filtering is absolutely not based on the thematic dimension of the profiles, and so it is not subject to the funnel effect. Another advantage of collaborative filtering is that the judgments of the users integrate not only a thematic dimension but also other factors relating to the quality of the documents such as diversity, the innovation, the adequacy with the public concerned, and so on.
Collaborative filtering is mainly based on the principle of communities. A community represents a set of users who share the same centers of interests. A collaborative filtering system exploits communities to generate the best recommendations. So the management of the communities (their creation, their evaluation, their perception) in collaborative filtering systems takes an important part because the quality of the recommendations sent to the users depends basically on the quality of the communities formed by the system. It exists different approaches to create communities. Most of the time the communities are generated from the computation of users proximity according to the history of their evaluations. The most popular approach to build the community of a user is that of the closest neighbors (Breese et al., 1998; Herlocker et al., 1999).
Many systems of recommendations are based partially or completely on filtering collaborative (Herlocker et al., 1999; Herlocker at al., 2004; Jin et al., 2004). The experiments carried out with these systems proved the effectiveness of collaborative filtering and the help brought to the users by the recommendations generated (usefulness of these recommendations). At all events, the positioning of new users within communities is not obvious and must be the subject of studies to come.

d. Others Collaborative Recommender Systems

Many other approaches propose alternative recommendation systems. For instance, *Siteseer* (Rücker & Polanco, 1997) or *Easy-DoR* (Chevalier et al., 2004) recommend documents to users according to the contents of their bookmark hierarchies. The main difference between these two approaches is that the second one takes into account the hierarchical organization of bookmarks.
Fab (Balabanovic & Shoham, 1997) or *Groupmark* (Pemberton et al., 2000) propose recommender systems based on groups of users, namely communities. *Fab* queries external search engines to find and recommend information to user groups built according to their information needs. *Groupmark*, recommends to users the groups which are adapted to him. A group can be considered as an information need. An alternative can be seen via *ReferalWeb* (Kautz et al., 1997) which relies on the word to mouth principle.

In addition to pull-based assistants, assistants have been developed to improve the second modality of the search process: browsing.

4. Browsing Assistants

The main lacks of browsing come from the cognitive effort that it implies. Indeed, the user does not know the hypertext structure and should remind which path he follows to reach a specific document: this commonly implies a cognitive overload. Moreover, the user has to identify which links are relevant to his information need. Nevertheless, when the number of links rises, this judgment is less easy. Navigation suffers mainly from problems coming from the architecture on which it is based, i.e. on the hypertext concept. These problems are cognitive overload and disorientation. For both of them, we present the approaches that aim at limiting these problems.

4.1. Individual Browsing Assistants

4.1.1 Cognitive Overload

In order to limit the cognitive effort induced by the hypertext, various tools were implemented to keep a trace of the visited documents.

a. Navigation "History" List. Navigation history is a basic browser feature pioneered by Mosaic in autumn 1993. Concretely, the browser automatically stores addresses of documents visualized by the user, in a chronological order of visit. Thus, the user can retrieve an encountered document at any time, provided the fact that he remembers the date of visit. This requirement makes the exploitation for a navigation history cognitively expensive. Traditionally, documents pertaining to a user's navigation history are represented by their title and URL in a widget list, as in Internet Explorer for example. However, such a representation is hard to exploit, in particular when searching for a document. As people may rather remember a document's presentation than its title, *Bookmap* (Hascoët, 2000) displays documents from the history by thumbnails, i.e. stamp sized pictures of each document. More precisely, in *Bookmap* such thumbnails are vertices of a graph whose edges reflect navigation paths. This system features in an integrated way:

- The visualization of the bookmarks hierarchies via a graph,
- The memorization of the documents the user has "to read,"
- A direct visualization of the navigation history list,
- A direct access to the most visited documents via their URL ("best of" documents).

This project is interesting because it uses "thumbnails" (reduced images of the documents) instead of traditional URLs or titles of the documents. Thanks to this hint, the user may visually have a direct outline of a document, which facilitates its evaluation. Unfortunately, the user must nevertheless provide a high cognitive effort during his navigation because the various documents are presented independently. Navigation visualization systems try to limit this issue.

b. Navigation Visualization. Navigation visualization systems can be considered as an evolution of the navigation history list. It makes it possible to graphically represent the visited documents as well as the possible links between these documents. We can give as example *WebMap* (Dömel, 1994). Thanks to such tools, the user can visualize not only the documents that he has previously visited but also the organization of these documents.

4.1.2. Disorientation

Disorientation is caused by the fact that the user loses the progression of his navigation. This is due to the fact that he does not understand the organization of the local hypertext anymore. To avoid that, cartographying the local hypertext has been proposed. Instead of only presenting the visited documents, these cartographies aim at presenting the documents visited within their local hypertext. Thus, the visited documents appear with the documents bound by hypertext links so that the user can have a global vision of the "local" hypertext in which he is located. As an example of cartography tools, we can give *Hyperspace* (Wood et al., 1995) or *Internet Cartographer* (http://www.inventix.com/).

Thanks to the various approaches suggested in the literature, the navigation task can be carried out under better conditions. But Information Retrieval on the Web also requires a research task. The following section presents the approaches aiming at improving the research task.

Letizia (Lieberman, 1995) uses the visited documents to initialize the user's profile. This latter is built starting from the terms that result from the visited documents and the actions of the user. During a document visit, *Letizia* downloads the contents of all the documents related to the active document in order to evaluate the matching with the user profile. The links towards relevant documents are presented to the user by means of an additional window.

These systems only tell the user potentially relevant documents for navigation in progress. However, it is also interesting to inform the user that the document he is visiting is relevant (or not relevant) for the topics that interests him. The *Syskill & Webert* (Pazzani, 1996) system aims at filling this task. The user can create as many topics as necessary. The topics must be organized in a non hierarchical list. Each time he finds a relevant (or non relevant) document for the topics he has created, the user announces it to the system. Then this one updates a profile corresponding to the topic. During navigation, the system indicates the user whether the document he visits is relevant (or not relevant) for his topics.

In a more general way, agents can also take into account the user's search environment (collection of applications) to deduce his needs and to recommend documents to him which are not necessarily connected (by hyperlinks) to the visited documents. *WBI* (Barrett et al., 1997), for example, studies the user's actions and organizes the visited documents into classes representing the user's needs. From these needs, *WBI* queries research tools to propose the search result to the users. It moreover proposes functionalities of history management, notification (announces the modifications in the contents of a document) and shortcuts generation (if the user daily follows the same sequence of bonds to reach a document everyday). This approach is based on a multi-agent architecture and the interface of *WBI* is integrated into Web documents.
The *Watson* system (Budzik & Hammond, 1999) innovates concerning the identification of regular expressions in the visited documents (as the addresses for example) in order to propose contextual services to the user. For an identified address, *Watson* generates an urban map allowing the user to locate it.
To limit user's disorientation, an alternative of such recommender systems can be seen in *Topic Maps* (Le Grand & Soto, 2005) which can produce a topic-based navigation of the document set.

4.2. Social Browsing Assistants

Social assistants which help a user to browse the Web can also be called browsing accelerators. They recommend the user links that the system considers as relevant for the user's navigation.
Webwatcher (Armstrong et al., 1995) used the user's information needs that can be considered as a query, to recommend the user relevant hypertext links. A user may judge each document he visits. For each relevant document, the system stores the user's relevance judgment. The system uses this information to identify links that are relevant for the user's information needs. To do this, the system prefetches the entire document that is linked to the visited one and use all stored information about each document associated with as well as the user's information needs to compute the relevance of each document. The hyperlinks corresponding to relevant document are indicated via specific icons added by the system within the visited document.
Broadway (Jaczynski & Trousse, 1997) is based on the same idea. It relies on the hypothesis that two users visiting the same documents in the same order implies that they are looking for the same information. The user profile corresponds to a user's navigation path associated with many behavior characteristics (e.g. reading time). The system stores all profiles to identify relevant information for a new user. To identify these documents, *Broadway* is based on a case-based reasoning algorithm that identifies the paths similar to a specific user's ones within store navigation. When similar navigation paths are found, *Broadway* recommends some documents found in retrieved navigations that the user will probably visit. It allows the user to accelerate his navigation by going directly to the required information.
Rather than recommending documents, some approaches (social navigation systems) offer users a shared space or a virtual world in which documents are displayed.
For instance, *FootPrints* (Wexelblat & Pattie, 1997) displays the browsing behavior of users in a 2D interface. Anyone can see others' navigation paths as well as documents that have been the most visited one…
CoBrow (Sidler et al., 1997) is a system that displays the hypertext structure in which visited documents occur. It also features communication functionalities between users visiting the same documents, thus transforming the Web into a virtual space where people meet on Web pages.
A more sophisticated system is illustrated by *StarWalker* (Chen, 1999) which offers a 3D virtual environment dedicated to a document collection browsing where users can interact, e.g. by exchanging information, communicating.

To help the user to improve his search process as a whole, he needs to be assisted during search related activities.

5. Improving Other Search Related Activities

5.1. Individual-based Improvement

Systems provide individual help to achieve both modalities of information search: browsing and querying. Moreover, activities such as document storing, document active reading, and document sharing are taken into account to improve users' experience and knowledge "capitalization." The following sections describe how

systems intend to mprove these activities.

5.1.1. Active Reading Individual Assistants

In past centuries, when books were rare and personal possessions, readers used to annotate and share their own copies (Jackson, 2002). Nowadays, annotating books borrowed from a library is considered defacement since they are public materials. However, some people still seek for the "dirtiest," i.e. most annotated copy available because they find previous readers' annotations valuable (Marshall, 1998). Now that documents are digitized, annotation value is reconsidered and introduced into the digital library through software called "annotation systems."

Currently, documents tend to be drawn up using word processing software and are mainly spread through networked computers. It is noticed that reading documents on-screen is less comfortable, slower and less persuasive than reading paper. Moreover, according to an experiment recounted by Sellen & Harper (2003), readers feel frustrated of not being able to annotate digital documents.

The need to annotate digital documents has been soon understood by researchers and companies that have been developing annotation systems since the early '1990s, consider for example *Commentor* (Röscheisen et al., 1994). Concretely, people consulting a digital document can select a passage, create an annotation using the appropriate software function, and then type in a comment. Once the annotation is created, it is generally displayed in context, as close as possible to its anchor, i.e. the selected passage like in *Amaya* (Kahan et al., 2002).

5.1.2. Document Management Individual Assistants

Nowadays, people have to manage huge amounts of digital documents that they receive or have searched for. Moreover, these digital documents can be multimedia, combining texts, pictures, sounds, and videos. With the advent of handy and user-friendly software, people can also easily produce their own documents. Furthermore, they are able to publish them on the Web so as others can retrieve them. Actually, organizing digital material is a key issue for being able to share and to find them later—as is it the case for paper documents. Jones et al. (2005) have observed that people prefer to organize documents themselves rather than to trust in a personal search engine. In fact, they feel that searching by querying makes them loosing control of their personal data, as they do not rely on the system's performance. Indeed, we present below three facilities for storing encountered documents.

a. Navigation History List. In Web context, browsers have soon allowed people to reach encountered documents again by keeping their navigation histories. Actually, the browser automatically fills the user's navigation history with every visited Webpage. However, a very restricted subset of these documents may really interests the user. That is the reason why browsers also provide a bookmarking feature that enables users to store interesting documents, in a more active way.

b. Personal Bookmarks. In paper document setting, "bookmarking a page" refers to the act of locating a specific page thanks to a thin marker, commonly made from paper or leather. Nowadays, people more and more work with digital documents; therefore bookmark facility has naturally been transposed from paper to the digital world. The browser's bookmarking feature enables readers to keep traces of an interesting documents by storing its URLs. Bookmarks are organized in a hierarchy similar to a file system: it is a tree whose nodes are folders and leaves are pointers to documents.

Concretely, a Web user that wants to keep a document for future use may insert it into his hierarchy. Back to 1998, users were already storing an average of three bookmarks per navigation session; experience showed that they became creating folders when their whole bookmarks do not fit anymore into the screen (Abrams et al., 1998). Actually, users mostly structure their bookmarks in an incremental fashion, corresponding to their use. Bookmarks are nowadays widely used as they are integrated to any browser. However, they suffer some limits that may be swept away by the use of annotations, as discussed in the following section.

c. Digital Annotations. Annotating a paper is an activity commonly practiced, mostly involved in active reading. On paper, an annotation is more informative than a bookmark for at least two reasons. First, the reader associates his annotation to a particular location within the document. This anchoring point usually represents the annotation's function, e.g. writing phrases in the header of a document may sum it up whereas crossing text out may express refutation. Secondly, an annotation may consist of a comment, expressed in the reader's own verbal representations that reflect his understanding of the document.

Such differences can also be noticed on digital documents. Indeed, software called "annotation systems" such

as *Amaya* (Kahan et al., 2002) provides annotation feature on documents of the Web or even on everything displayable—users annotate screen shots—thanks to *ScreenCrayons* (Olsen et al., 2004). Once created, annotations can be stored in the user's personal annotation repository, which is usually a hierarchy. Later on, a user can directly retrieve a particular annotation and its related document, by browsing his annotation hierarchy, or by querying the system with keywords.

5.2. Social-based Improvement

Nowadays, modern systems do not consider that users act alone any longer. Indeed, they are viewed as community members that can benefit from an identified group, and *vice versa*. Systems that are aware of such links between people can improve users' activities by providing specific tools. In this section, we describe how activities identified in section 1.3 are improved this way.

5.2.1. Active Reading Social Assistants

Annotation systems enable readers to formulate annotations such as comments about Webpage contents. Annotations are attached to specific locations called "anchoring points," e.g. a word, a paragraph. When they are stored on a dedicated annotation server—as in *Amaya* (Kahan et al., 2002) or *Pharos* (Bouthors & Dedieu, 2000) for example—they can be retrieved along with documents. Regarding privacy, users can specify annotations' visibility, e.g. private, public, and restricted to some defined groups. Marshall & Brush (2004) have shown that public annotations are private ones at first. In fact, annotators have to reformulate them before their publication. Indeed, provided that they have appropriate grants, readers can both read documents and associated annotations.

In their slightest form, annotations only consist in highlighted passages of documents. Considering that these passages reflect *n* users' interests, Marshall (1998) defines the "*n*-way consensus": it is a new document obtained by extracting the passages commonly highlighted by almost *n* readers. By relaxing the constraint on *n*, a user can progressively view passages that have been highlighted by less and less people. Therefore, users' annotations can be useful for identifying passages that have be judged as important by previous readers.

Considering a collective use, Wolfe and Neuwirth (2001) observed that annotations allow readers to provide feedback to writers or promote communication with collaborators as well as to address remarks directed to future readers while authoring. Now, one may consider a comment of an annotation as subjective because it expresses the annotator's point of view. In order to allow readers to discuss documents in context, systems treat annotations as annotatable contents. Therefore, an annotation system can support asynchronous discussions within "discussion threads." A discussion thread is a hierarchy of annotations that is ordered by their timestamps, and rooted to its anchor, i.e. a document.

Concretely, argumentative discussion can take place in the context of a document, thanks to argumentative annotations. Such annotations reflect readers' opinions such as gradual confirmation or refutation. For example, annotations in the *Collate* project have been used by scientists for collaborative interpretation and indexation of movies (Thiel et al., 2004). Since traditional visualization of annotations—by a specific icon—can clutter widely annotated documents, Cabanac et al. (2005) propose to evaluate the consensus of discussion threads for adapting their visualization. Thus, users can ask for visually emphasize confirmed annotations rather than refuted ones, for example. Therefore, they can focus on ongoing conversations or on confirmed propositions that have been validated by people.

5.2.2. Document Management Social Assistants

Regarding personal bookmarks as a collective source of information has early been proposed by Keller et al. (1997) with the *WebTagger* system, for example.

A new way for storing encountered documents is called "social bookmarking," as it also enables users to share them on the Internet. Indeed, we currently attend the growth of such numerous online services (Hammond et al., 2005) such as *OneView* (http://beta.oneview.de), the very popular *del.icio.us* (http://del.icio.us) or *Flickr* (http://www.flickr.com), for instance. Such services enable Internet users to store online documents' URLs along with a comment and one or more *tags*. A "tag" is a free-form text of the user choosing; it aims at describing the document contents, e.g. "chapter" and "social IR" may be used for describing this chapter.

In traditional library settings, indexing process consists in identifying a resource with a **proper** call number based on content. In contrast with this, social bookmarking users produce **numerous** labels for each resource,

which can be content-based, usage-based, etc. The "social" aspect of these systems allows people to view *tags* associated to a particular document; they can also view every document associated to a given *tag*. For example, one can retrieve each document that people tagged with "computer science" via the `http://del.icio.us/tag/computer+science` URL. In fact, one can find interesting documents by browsing users' tags.

5.2.3. Social Document Sharing Assistants

Social bookmarking is a simple but effective way for sharing documents. In order to improve the efficiency of this approach, systems encourage users to reuse previously defined *tags*. Moreover, Golder & Huberman (2006) show that people tend to uniform names of the tags they create, by employing singular words rather than plurals, for example. They also learn from others' naming patterns, leading them to employ tags that are often used, e.g. "web2.0" *vs* "The Web 2." As a consequence, this kind of "tag convergence" helps users to find nuggets of information more easily.

"Collective intelligence" is a key concept of Web 2.0 that relies on information sharing within a community. However, general social bookmarking services are too open for that; in order to gather contributions related to specific topics, community-based services have been needed. The most famous example may be the *Connotea* service (http://www.connotea.org) provided by the *Nature Publishing Group* since 2005. It has been designed for sharing scientific references between researchers (Lund et al., 2005). For an industrial context, the *DogEar* system (Millen et al., 2006) allows employees at I.B.M. to tag and share resources, helping them to find experts on certain domains.

As seen before, more and more related activities are being developed in the field of information searching to help users but most of the issues still concern future trends and need to be experimented.

6. Future Trends

Information seeking is based on many activities requiring many capabilities to be effective. Information Retrieval Systems are the main and the most powerful tools to find a needle in the haystack that corresponds to the Web. Indeed, the volume, the structure, the granularity, the validity of available information on the Web cannot be manually handled. However, we can notice that many approaches have been provided to help a user during his search process. In the Web 2.0 context, collaborative tools that benefit from other users' experience progressively replace individual search assistants. The aim of the current Web 2.0 is to gather applications, services, and collective intelligence. While this evolution is not totally mature and widespread enough—one may describe it as a "beta" version—the next step of Web evolution is, according to Markoff (2006), already envisioned: the Web 3.0. New generation systems may harness semantic (such as ontologies) for understanding data, logics for deducing facts, users' profiles for personalized reasoning, extracting accurate data rather than documents from the Web … This evolution joins the Semantic Web concepts mentioned since 1999 by Tim Berners-Lee, the creator of the original World Wide Web.

Unfortunately, most of search assistants are not enough spread on the Web or even not accepted by users. Indeed, the information seeking process is split into multiple tools that do not communicate together, e.g. a browsing accelerator, an external annotation system. Thus, users—who may not care for using many different tools together to find relevant information—are only interested in tools that quickly return a result, even if it is approximate or incomplete or sometimes non relevant. However, the sole use of an Information Retrieval System without associating related activities like active reading, memorization… may not be really effective because incomplete.

Therefore, integrated search assistants should not only mix query and browsing features but also incorporate annotating, memorizing, and sharing features. This limit is not really solved in the Web 2.0 context. Even if the Web is nowadays supporting more and more social or collective applications, the whole of activities in the search process are still independent and treated separately. The Web 2.0 evolution has also some limits. Social bookmarking systems for instance are based on users' active participation. Such participatory systems suffer from "free-riding" behaviors: when people only benefit from systems without really contributing to them. In order to limit this issue, observations of social psychology studies may be considered. For example, Ling et al. (2005) show that a person contributes more when he is aware that his contributions *i*) are judged useful by others and *ii*) are special, i.e. nobody else could have contributed that way. Moreover Wu et al. (2006) underlines some limits of folksonomies like the need to organize them when their number rises… One

interesting drawback of such approaches is also the trend to uniformity. Indeed, to share information or tags, users have to use intelligible words and so, use the same terms as other users. For instance, a document may be tagged by the term "database" and not "db." This implies that systems should respect the user's point of view while providing powerful sharing capabilities. As a solution, such systems should manage two distinct user's "views" in order to respect these two aspects: a personal view and a social view. The personal view may be limited to the user who should be allowed to manage it as he likes.

So, social approaches should not limit the user to any "social" correct way of thinking: a kind of single thought. Any user has its own "right to think which is shared with other in the same or in a different form. Moreover social approaches may motivate users to participate (Beenen et al., 2004) and may lead to limit "free-riding" penalizing behaviors.

Despite these drawbacks, social information retrieval must be developed to allow anyone to benefit from other users' experience. For instance, newbies should profit and learn from experts' experience. The modality is not really defined and everything is possible. To do that, users and their behavior have to be more detailed in users' profiles that represent the basis of search assistants, being either individual or collective. On top of the accurate description of users along with their needs, profiles may also integrate users' constrains. Indeed, the future of search assistants may be based on an improved adaptability of the search process. The adaptability may be seen at multiple levels. For example, the assistant must be task-oriented and must take into account the time the user can spend for his search process. Moreover, it should adapt retrieved information to user's incapacities, language constraints, format constraints… Thus, in addition to be retrieved, relevant information will have to be adapted, transformed (e.g. from Microsoft Word's format to Adobe's PDF one), and extracted to be really and efficiently exploited by the user. A first solution to such an adaptation may be seen through the widespread of Web services or more generally services specifically selected in addition to the search engine to complete its capabilities. It may adapt the search process itself or the results to a specific user. As a conclusion, adaptation methods associated with the Semantic Web concepts could provide an open architecture, through which, for instance, one could offer and describe services in the same way as it is done today for information.

7. References

Abrams, D., Baecker, R., & Chignell, M., 1998. Information Archiving with Bookmarks: Personal Web Space Construction and Organization. In *CHI '98: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 41-48, New York, NY, USA. ACM Press/A-W Publishing Co.

Adler, M. J., & van Doren, C., 1972. *How to Read a Book*. Simon & Shuster, NY.

Agosti, M. (1996). An Overview of Hypertext. In (Agosti & Smeaton, 1996), chapter 2,pages 27–47.

Agosti, M., & Ferro, N., 2005. Annotations as Context for Searching Documents. In *CoLIS '05: Proceedings of the 5th International Conference on Conceptions of Library and Information Sciences*, volume 3507 of *LNCS*, pages 155-170. Springer.

Agosti, M., & Melucci, M., 2000. Information Retrieval on the Web. M. *Agosti, F. Crestani & G. Pasi Eds., ESSIR 2000, Lecture Notes in Computer Science 1980*, Springer Verlag (Berlin), pages. 242–285.

Agosti, M. and Smeaton, A. F. (1996). *Information Retrieval and Hypertext*. Kluwer Academic Publishers, Dordrecht.

Armstrong, R., Freitag, D., & Joachims, T., 1995. Webwatcher: machine learning and hypertext, In *Proceedings of the 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, Stanford University, California, USA.

Baeza-Yates, R., & Ribeiro-Neto, B., 1999. Modern Information Retrieval, First edition, Addison Wesley

Balabanovic, M., & Shoham, Y., 1997. Fab: content-based, collaborative recommendations. *Communications of the ACM*, 40(3):66 –72.

Barrett, R., Maglio, P.P., & Kellem D.C., 1997. How to personalize the web, *International ACM Conference on Human Factors and Computing Systems (CHI)*, Atlanta Georgia, pp 75-82.

Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P., & Kraut, R. E. (2004). Using Social Psychology to Motivate Contributions to Online Communities. In *CSCW '04: Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 212–221, New York, NY, USA. ACM Press.

Benford, S., Snowdon, D, Greenhalgh, C., Knox I., & Brown C., 1995. VR-Vibe: a virtual environment for cooperative information retrieval, *EuroGraphics*, 14(3): 349–360.

Bergman, M. K., 2001, The Deep web: surfacing the hidden value, *Journal of Electronic Publishing from the University of Michigan,* Retrieved January 10, 2007, http://www.press.umich.edu/jep/07-01/bergman.html.

Belkin, N. J., & Croft, W. B., 1992. Information Filtering and Information Retrieval: Two sides of the Same Coin? *Communication of the ACM*, 35(12), 29-38.

Bharat, K., & Henzinger, M., 1998. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Distributed Retrieval*, pp. 104-111.

Bottraud, J. C., Bisson, G., & Bruandet, M. F., 2003 An adaptive information research personnal assistant. In *proceedings of the workshop AI2IA (Artificial Intelligence, Information Access and Mobile Computing) IJCAI*.

Bouthors, V., & Dedieu, O., 1999. Pharos, a cooperative infrastructure for web knowledge sharing. Technical report num. 3679, ISSN 0249-6399, Institut de Recherche en Informatique et en Automatique (INRIA).

Breese, J.S., Heckerman, D., & Kadie, C., 1998. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of 14th conference on Uncertainty in Artificial Intelligence* (UAI'98), Wisconsin (US), pp. 43-52.

Brin, S., & Page, L., 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International World Wide Web Conference* (WWW7), pp. 107-117.

Budzik, J., & Hammond, K., 1999. Watson : anticipating and contextualizing information needs, *Annual Meeting of the American Society for Information Science (ASIS)*, Washington.

Cabanac, G., Chevalier, M., Chrisment, C., & Julien, C., 2005. A Social Validation of Collaborative Annotations on Digital Documents. In Boujut, J.-F., editor, *International Workshop on Annotation for Collaboration*, pages 31-40, Paris. Programme société de l'information, CNRS.

Carrière, J., & Kazman, R., 1997. WebQuery : Searching and visualizing the web through connectivity, In *Proceedings of the 6th International World Wide Web Conference* (WWW6), pp. 701-711.

Chen, C., 1999. The StarWalker virtual environment - an integrative design for social navigation. *Human-Computer Interaction: Communication, Cooperation, & Application Design, Proceedings of HCI International'99, the 8th International Conference on Human-Computer Interaction*, Volume 2, pp. 207-211.

Chevalier, M., Chrisment, C., & Julien, C., 2004. Helping people searching the web: towards an adaptive and a social system. *IADIS/WWW Internet 2004*, Pedro Asaias, Nitya Karmakar (Eds.): IADIS.

Cho, Y. H., Kyeong, J., & Kim, S. H., 2002. A personalized recommender system based on web usage mining and decision tree induction, *Expert System with Applications*, 23(3):329 –342.

Cutting, D. R., Karger, D. R., & Pederson, J. O., 1993. Constant interaction-time scatter/gather browsing of very large document collections, In *14th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp 121-131.

Dömel, P., 1994. Webmap - a graphical hypertext navigation tool, *2nd International World Wide Web Conference*, pp. 785-798

Dreilinger, D., & Howe A. E. 1997. Experiences with selecting search engines using meta-search, *ACM Transactions on Information Systems*, 15(3):195 –222.

Fidel, R., Bruce, H., Pejtersen, A. M., Dumais, S., Grudin, J., & Poltrock, S., 2000. Collaborative Information Retrieval (CIR). *The New Review of Information Behaviour Research*, pp. 235-247.

Fraenkel, A. S. and Klein, S. T. (1999). Information Retrieval from Annotated Texts. *J. Am. Soc. Inf. Sci.*, 50(10):845-854.

Frommholz, I. and Fuhr, N., 2006. Probabilistic, Object-oriented Logics for Annotation-based Retrieval in Digital Libraries. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 55-64, New York, NY, USA. ACM Press.

Garfield, E., 1972. Citation analysis as a tool in journal evaluation. *Science*, 178(4060), pp. 471-479. www.garfield.library.upenn.edu/essays/V1p527y1962-73.pdf.

Godoy D., & Amandi A., 2006. Modeling user interests by conceptual clustering, *Information Systems*, 31(4):247–265.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D., 1992. Using collaborative filtering to weave an information tapestry. *Communication of the ACM, Information filtering*, 35(12):61–70.

Golder, S. A. and Huberman, B. A., 2006. Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science*, 32(2):198-208.

Good, N., Schafer, J., Konstan, J., Borchers, A., Sarwar, B., Herlocker, J., & Riedl, J., 1999. Combining collaborative filtering with personnal agents for better recommendations. In *Proceeding of AAAI, AAAI Press*, 35:439-446.

Gravano L., Garcia-Molina H., & Tomasic A., 1999. GloSS: text-source discovery over the internet, ACM Transactions on Database Systems, 24(2):229-264.

GVU, 1998. 10th WWW User Survey, Graphic, visualisation & usability center (GVU), 1998. Retrieved January 10, 2007, http://www.gvu.gatech.edu/user_surveys/survey-1998-10/.

Hammond, T., Hannay, T., Lund, B., & Scott, J., 2005. Social Bookmarking Tools (I): A General Review. *D-Lib Magazine*, 11(4).

Haase, P., Stojanovic, N., Sure, Y., & Völker, J., 2005. On Personalized Information Retrieval in Semantics-Based Peer-to-Peer Systems. In *Proceedings of the BTW-Workshop "WebDB Meets IR"*. Retrieved January 10, 2007, http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/2005_webdbir.pdf

Hansen, P., & Järvelin, K., 2005. Collaborative Information Retrieval in an information-intensive domain. *Information Processing & Management Journal, Elsevier Science*, 41(2005):1101 –1119.

Hascoët, M., 2000. A user interface combining navigation aids, In *Proceedings of the 11th International ACM Hypertext Conference*, pp 224-225.

Hauffe, H., 1994. Is citation analysis a tool for evaluation of scientific contributions ?, In *13th Winterworkshop on Biochemical and Clinical Aspects of Pteridines*. Retrieved January 10, 2007, http://www.uibk.ac.at/ub/ueber_uns/publikationen/hauffe_is_citation_analysis_a_tool.html

Herlocker, J.L., Konstan, A.J., Borchers, A., & Riedl, J., 1999. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd International ACM Conference on Research and Development in Information Retrieval* (SIGIR'99), pp. 230-237.

Herlocker, J.L., Konstan, A.J., Terveen, L., & Riedl, J., 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems* (TOIS), 22 (1), pp. 5-53.

Hetzler, B., Harris, W. M., Havre, S., & Whitney P., 1998. Visualizing the full spectrum of document relationships, In *5th International Conference of the International Society for Knowledge Organization (ISKO)*, pages 168–175.

Hölscher, C. and Strube, G., 2000. Web Search Behavior of Internet Experts and Newbies. In *Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications networking*, pages 337-346, Amsterdam, The Netherlands, The Netherlands. North-Holland Publishing Co.

Ide, E., 1971. New Experiments in Information Retrieval., In *G. Salton Editor, The SMART Retrieval System – Experiments in Automatic Document Processing*, Prentice Hall Inc., Englewood Cliffs (NJ), pages 337–354.

Jackson, H. J., 2002. *Marginalia: Readers Writing in Books*. Yale University Press.

Jaczynski M., & Trousse B., 1997. Broadway: a world wide web browsing advisor reusing past navigations from a group of users, In *Proceedings of the 3rd UK Case-Based Reasoning Workshop (UKCBR'97)*.

Jansen, B. J., Spink, A., & Saracevic. T. Real life, real users and reals needs: A study and analysis of users queries on the Web. 2000. *Information Processing & Management Journal, Elsevier Science*, 36(2):207–227.

Jin, H., Ning, X., & Chen, H. 2006. Efficient search for peer-to-peer information retrieval using semantic small world. In *Proceedings of the 15th International Conference on World Wide Web*. ACM Press, New York, NY, 1003-1004.

Jin, R., Chai, J.Y., & Si, L., 2004. An automatic weighting scheme for collaborative filtering. In *Proceedings of the 27th International ACM Conference on Research and Development in Information Retrieval*. SIGIR'04, pp. 337-344.

Jones, W., Phuwanartnurak, A. J., Gill, R., & Bruce, H., 2005. Don't Take My Folders Away!: Organizing Personal Information to Get Things Done. In *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, pages 1505-1508, New York, NY, USA. ACM Press.

Kahan, J., Koivunen, M.-R., Prud'Hommeaux, E., & Swick, R. R., 2002. Annotea: an open rdf infrastructure for shared web annotations. *Computer Networks*, 32(5):589-608.

Karamuftuoglu, M., 1998. Collaborative information retrieval: Towards a social informatics view of IR interaction. *Journal of the American Society for Information Science*, 49(12):1070–1080.

Kautz, H., Selman, B., & Shah, M. (1997). ReferralWeb: combining social networks and collaborative filtering. *Commun. ACM*, 40(3):63–65.

Kaye, J. J., Vertesi, J., Avery, S., Dafoe, A., David, S., Onaga, L., Rosero, I., & Pinch, T., 2006. To Have and to Hold: Exploring the Personal Archive. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 275-284, New York, NY, USA. ACM Press.

Keller, R. M., Wolfe, S. R., Chen, J. R., Rabinowitz, J. L., & Mathe, N., 1997. A Bookmarking Service for Organizing and Sharing URLs. In *Selected papers from the 6th international conference on World Wide Web*, pages 1103-1114, Essex, UK. Elsevier Science Publishers Ltd.

Kleinberg, J. M., 1998. Authoritative sources in a hyperlinked environment. In *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 668-677.

Klemm, F., & Aberer, K., 2005. Aggregation of a Term Vocabulary for Peer-to-Peer Information Retrieval: a DHT Stress Test, In *Proceedings of the Third International Workshop on Databases, Information Systems and Peer-to-Peer Computing* (DBISP2P 2005).

Konstan, J. A., Miller, B. N., Maltz, D. Herlocker, J. L.,. Gordon L. R., & Riedl J., 1997. Grouplens: Applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3):77–87.

Korfhage, R. R., 1997. *Information storage and retrieval*. Wiley computer publishing.

Krulwich, B., 1997. Lifestyle finder: Intelligent user profiling using large-scale demographic data. *AI Magazine*, 18(2):37-45.

Kwok, S. H, 2006. P2P searching trends: 2002-2004. *Information Management and Processing Journal*, Elsevier Science, Vol. 42(1), pp. 237-247.

Le Grand, B., & Soto, M., 2005. Topic Maps, RDF Graphs and Ontologies Visualization, in Visualizing the Semantic Web. Ed. Geroimenko V. & Chen C. Springer., 2nd edition.

Leouski, A. V., & Croft, W. B., 1996. An Evaluation of Techniques for Clustering Search Results, *Technical Report IR-76*.

Lieberman, H., 2002. Letizia: An agent that assists web browsing. *In Proceedings of the IJCAI*, pp. 924-929, 2002.

Ling, K., Beenen, G., Ludford, P., Wang, X., Chang, K., Li, X., Cosley, D., Frankowski, D., Rashid, L. T. A. M., Resnick, P., & Kraut, R., 2005. Using Social Psychology to Motivate Contributions to Online Communities. *Journal of Computer-Mediated Communication*, 10(4):10.

Liu, M., 1993. The complexities of citation practice: a review of citation studies. *Journal of Documentation*, 49(4):370–408.

Lund, B., Hammond, T., Flack, M., & Hannay, T., 2005. Social Bookmarking Tools (II): A Case Study - Connotea. *D-Lib Magazine*, 11(4).

Maglio P. P., Barrett R., Campbell C. S., & Selker T., 2000. SUITOR: an attentive information system, *International ACM Conference on Intelligent User Interfaces (IUI)*, pp 169-176.

Malone, T. W., Grant, K. R., Turbak, F. A., Brobst S. A., & Cohen M. D., 1987. Intelligent Information Sharing Systems. *Communications of the ACM*, 30(5):390–402.

Markoff, J., 2006. Entrepreneurs See a Web Guided by Common Sense, *The New-York Times*, November 12, 2006.

Marshall, C. C., 1998. Toward an ecology of hypertext annotation. In *HYPERTEXT '98: Proceedings of the 9th ACM conference on Hypertext and hypermedia*, pages 40-49, New York, NY, USA. ACM Press.

Marshall, C. C. and Brush, A. J. B. (2004). Exploring the relationship between personal and public annotations. In *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 349–357, New York, NY, USA. ACM Press.

Mechkour, M., Harper, D. J., & Muresan G., 1998. The webcluster project: using clustering for mediating access to the world wide web », In *Proceedings of the 21st International ACM SIGIR Conference on Research and development in Information Retrieval*, pp 357-358, August 24-28, 1998.

Memmi D., & Nérot O., 2003. Building virtual communitites for information retrieval, in Favela & D. Decouchant (eds), *GroupWare: Design, Implementation and Use,* Springer, Berlin.

Millen, D. R., Feinberg, J., & Kerr, B., 2006. Dogear: Social Bookmarking in the Enterprise. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 111-120, New York, NY, USA. ACM Press.

Milojicic, D. S., Kalogeraki, V., Lukose, R., Nagaraja, K., Pruyne, J., Richard, B., Rollins, S., & Xu., Z., 2002. Peer-to-peer computing. Technical Report HPL-2002-57, HP Lab.

Montaner, M., Lopez, B., & Rosa, J. L. D. L., 2003. A taxonomy of recommender agents on the internet. *Artificial Intelligence Review*, vol. 19, pp. 285-330.

Olsen, D. R. J., Taufer, T., & Fails, J. A., 2004. ScreenCrayons: Annotating Anything. In *UIST '04: Proceedings of the 17th annual ACM symposium on User Interface Software and Technology*, pages 165-174, New York, NY, USA. ACM Press.

Pazzani, M., 1999. A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, 13(5-6), pp. 393-408.

Pazzani, M., Muramatsu, J. and Billsus, D., 1996. Syskill&Webert: Identifying interesting web sites. *In proceedings of the thirteenth National Conference on Artificial Intelligence,* pp. 54-61.

Pejtersen, A. M. and Fidel, R., 1998. A framework for centred evaluation and design: a case study of information retrieval on the web. Working paper for mira workshop, Grenoble, France. Retrieved January 10, 2007, http://www.dcs.gla.ac.uk/mira/workshops/nancy/amprf.html.

Pemberton, D., Rodden, T., & Procter, R., 2000. GroupMark: a WWW recommander system combining coopérative and information filtering, In *Proceedings of the 6th ERCIM Workshop* "User Interfaces for All".

Prekop, P., 2002. A qualitative study of collaborative information seeking, *Journal of Documentation*, 58(5):533–547.

Price, M. N., Schilit, B. N., & Golovchinsky, G., 1998. Xlibris: The active Reading Machine. *In CHI'98 conference summary on Human factors in computing systems*, pp. 22-23, ACM Press, 1998.

Rocchio, J. J., 1971. Relevance Feedback in Information Retrieval, In *G. Salton Editor, The SMART Retrieval System - Experiments in Automatic Document Processing*, Prentice Hall Inc., Englewood Cliffs (NJ), pp. 313-323.

Röscheisen, M., Mogensen, C., & Winograd, T., 1994. Shared web annotations as a platform for third-party value-added, information providers : Architecture, protocols, & usage examples. Technical report CSDTR/DLTR, Stanford, CA, USA.

Rücker, J., & Polanco, M. J., 1997. Siteseer: Personalized navigation for the web. *Communications of the ACM*, 40(3):73–75.

Salton, G., & Buckley, C., 1990. Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science*, 41(4):288–297.

Samulowitz M., Michahelles F., & Linhoff-Popien C., 2001. Capeus: An architecture for context-aware selection and execution of services. In *new developments in distributed applications and interoperable systems*, pp. 23–39.

Sannomiya, T., Amagasa, T., Yoshikawa, M., & Uemura, S. (2001). A Framework for Sharing Personal Annotations on Web Resources Using XML. In *ITVE '01: Proceedings of the workshop on Information technology for virtual enterprises*, pages 40-48, Washington, DC, USA. IEEE Computer Society.

Savoy, J., & Picard, J., 2001. Retrieval effectiveness on the web. *Information Processing & Management Journal, Elsevier Science*, 37(4):543–569.

Sellen, A. J. and Harper, R. H., 2003. *The Myth of the Paperless Office*. MIT Press, Cambridge, MA, USA.

Shapira, B., Shoval, P., & Hanani, U., 1997. Stereotypes In Information Filtering Systems. *Information Processing & Management*, 33(3)273–287.

Shneiderman, B., 1998. Designing the User Interface, Addison-Wesley.

Sidler, G., Scott A., & Wolf H., 1997. Collaborative Browsing in the World Wide Web. In *Proceedings of the 8th Joint European Networking Conference*.

Sonnenwald, D., & Pierce, L. G., 2000. Information behaviour in dynamic group work contexts: interwoven situational awareness dense social networks and contested collaboration in command and control, *Information Processing and Managemnt.* 36(3):461–479.

Talja, S., 2002. Information sharing in academic communities: Types and levels of collaboration in information seeking and use. *New Review of Information Behaviour Research*. 3, pp. 143-159.

Teevan, J., Dumais, S. T. & Horvitz, E., 2005. Personalizing Search via Automated Analysis of Interests and Activities. *In proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval*, pp. 449-456, ACM Press.

Thiel, U., Brocks, H., Frommholz, I., Dirsch-Weigand, A., Keiper, J., Stein, A., & Neuhold, E. J., 2004. COLLATE - A collaboratory supporting research on historic European films. *Int. J. Digit. Libr.*, 4(1):8-12.

Wexelblat, A., & Pattie M., 1999. Footprints: History-Rich Tools for Information Foraging. ACM SIGCHI Conference on Human Factors in Computing Systems, In *CHI'99 Proceedings*, ACM Press.

Wolfe, J. L. & Neuwirth, C. M., 2001. From the Margins to the Center - The Future of Annotation. *Journal of Business and Technical Communication*, 15(3):333–371.

Wood, A., Drew, N., Beale, R., & Hendley, B., 1995. Hyperspace: web browsing with visualisation, In *Proceedings of the 3rd International World Wide Web Conference (WWW3)*, pp. 21-25.

Wu, J., & Aberer, K., 2004. Using SiteRank for decentralized computation of Web document ranking, In *Adaptive Hypermedia and adaptive Web-based Systems*, LNCS 3137: 265-274.

Wu, X., Zhang, L., & Yu, Y. (2006). Exploring Social Annotations for the Semantic Web. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 417–426, New York, NY, USA. ACM Press.

Yager, R. R., 2002. Fuzzy logic methods in recommender systems. *Fuzzy sets and Systems*. Elsevier, 136(2):133–149.

Yang, B., & Garcia-Molina, H., 2002. Improving search in peer-to-peer networks, In *Proceedings of the 22nd international conference on distributed computing systems*, pp. 5-14.

Zamir, O., 1998. Visualisation of search results in document retrieval systems, General Examination, University of Washington, 1998.