HELPING PEOPLE SEARCHING THE WEB: TOWARDS AN ADAPTIVE AND A SOCIAL SYSTEM

Max Chevalier
IRIT/SIG, 118 route de Narbonne, F-31062 Toulouse cedex 04
LGC, 129 avenue de Rangueil, F-31077 Toulouse cedex

Claude Chrisment
IRIT/SIG, 118 route de Narbonne, F-31062 Toulouse cedex 04

Christine Julien
IRIT/SIG, 118 route de Narbonne, F-31062 Toulouse cedex 04

ABSTRACT
Searching information on the Web requires patience and perseverance. A user who is looking for some information should be lucky or experimented... The traditional way to search relevant information on the Web is to use a search engine. A search engine answers a user’s query which represents his information needs with (relevant) documents. But a Web information seeking process is not only a retrieval task. It also includes a browsing task, which is, not really taken into account today. Furthermore, search tools cannot ignore the user’s skills/knowledge because the query and so the search results directly depend on it. However, the Web, a representative example of the communication era, is not really a collaborative medium. Indeed, most of the Web users are individualists and do not share any information with others. Due to these limitations, a state of the art of user-oriented systems which try to facilitate the search task is done. This state of the art is also taking into account collective approaches that can be implemented to share information among users having the same point of interest.

In this context, the proposed approach is user-oriented and is based on an information share through which a person automatically gets information from others thanks to his information needs. It is aimed at helping the user while he is searching the Web. This approach has been implemented in a system called Easy-DOR.

KEYWORDS

1. INTRODUCTION
Nowadays the World Wide Web (or shortly Web) is the most privileged information source with its billions of Web documents. To identify those that are relevant to his information need, the user is using search engines or/and browsing. But, even if he uses such tools, the Net surfer is frequently facing the Web alone without any external assistance. So, a “social”1 and adaptive approach that helps him during his information seeking process becomes a necessity. The proposed one relies on an adaptive system which can be considered as a friend that helps the user in suggesting him some pieces of relevant information. The latter come from other surfers sharing the same point of interest. The proposition emerged from a study of the search task and particularly the user implication in such a process. This paper is organized as follows: the second section presents the Web information retrieval process and its limits. The third section presents the way a user can be assisted during his information seeking task. This section underlines either general approaches or social ones. The fourth section presents the proposed approach and its implementation.

1 The term “social” is used to characterize systems that share information between users without an obligatory human intervening.
2. FINDING RELEVANT INFORMATION ON THE WEB

Nowadays, the Web represents the most important information source with its billion of Web pages and documents. The main characteristics of the Web are that information is distributed, is not “stable” (created, modified or suppressed) and not persistent (Web servers are frequently down). To find relevant information to his needs, a user jointly uses search engines and browsing. To obtain the best results possible, the Net surfer should have a good knowledge of the queried domain and a good “practical” knowledge (Hölscher, 2000).

The knowledge of the queried domain is the most important required skill/knowledge in the information seeking task because the analysis of a document content and the query formulation are directly depending on it. This knowledge corresponds to the expertise the user has in a specific information field.

The “practical” knowledge is composed of skills that are necessary to surf the Web: how to browse the Web ?, how to query a search engine ?, how to combine these two methods (browsing and searching) ?, what are the URLs of relevant search engines ?...

But even if the user has a good knowledge, searching information on the Web suffers from some limitations. First of all, the structure of the Web itself implies that information related to the same topic is randomly located on the Web. Furthermore, no complete index exists to allow the user to precisely locate relevant information. So, the user has to be assisted to find relevant information in the huge amount of available documents. Search engines have been developed in this way but considering the retrieval achievement, each search engine is containing a different document index and is limited to only a part of the Web available documents (less than 20% of the visible Web). Fortunately, the issues encountered when finding relevant information on the Web are not limited to search engines. Some issues can also be identified during the browsing task. The hypertext structure of the Web implies a cognitive effort for the Net surfer to model the local organization of web documents and to avoid being lost in the hypertext.

Besides these practical limitations the Web also suffers from a poor social characteristic. Oddly, in the communication civilization, Net surfers are dramatically keeping information for themselves without sharing it with potentially interested people. This poor social characteristic makes the information seeking process on the Web solitary. So, considering all these limits, searching information on the Web requires patience, perseverance and sometimes luck. It requires methods and systems to make the Net surfer’s seeking task easier and/or more effective.

3. HELPING USERS SEEKING INFORMATION ON THE WEB

This section underlines the main approaches used to help the user during his information seeking process. It first presents the main solutions applied to the information retrieval task and then those which are related to the browsing task. General approaches are presented as well as social ones.

3.1 Helping Users Retrieving Information

3.1.1 General Approaches

The first issue when retrieving information is the query formulation. Indeed, according to his domain knowledge, the user can encounter some issues when transforming his mental representation of his needs into a formal one (a list of keywords). To limit these issues, most search engines offer categories where documents that are indexed are clustered. Through these categories, the user can interactively search documents that are relevant to his needs. In the same way, (Harper, 1999), through WebCluster, proposes a mediated access to indexed documents. After formulating the query, the selection of the search engine (information source) can be an issue. Indeed, retrieved documents quality depend on the selected search engine since each search engine indexes different documents. So, through the amount of general, domain specific search engines, the Net surfer has to select the most accurate and appropriate one according to his information needs. A solution has been proposed in GloSS (Gravano, 1999) which builds an index of several information sources and then allows a user to select an appropriate source according to his needs (query). As
a result, a set of relevant documents is presented to the user. Sometimes, the user may not formulate the query in the most accurate form, which implies that some retrieved documents are not relevant. Some search engines like Mercure (Boughanem, 1999) are featuring query reformulation (via a relevance feedback system for instance). To improve the query formulation some approaches use the user context (Bottraud, 2003). In spite of these effective retrieving tools, the main issue of the retrieval task is still the management of retrieved documents. The common way to present the retrieved documents to the user is the ranked list. All the documents are presented within a list that is sorted by relevance (according to the system). This presentation is easy to understand, easy to construct but it is only adapted to a low number of results (Cugini, 2000). This visualization is under-informative and is not appropriate for long list (>20 documents) of documents because no relation is visible between documents within the list. The relation between the query and the documents content (even if queried terms appear underlined in the snippet) are also invisible. Furthermore, to evaluate the real relevance of a retrieved documents, the user has to browse all the retrieved documents content. This task remains a too big cognitive effort for users who commonly visit the first twenty or thirty documents (Spink, 2002). To solve this issue and to allow users to manage search results (even if it is composed of thousands of documents) many re-ranking techniques and visualization interfaces have been proposed. The re-ranking techniques organize retrieved documents thanks to a user’s preferences. Visualization techniques use textual cluster, 2D or 3D metaphors principally aimed at presenting the whole search results at a glance. A survey of such visualization techniques can be found in (Zamir, 1998). Unfortunately, querying only one search engine implies that the result is only a sub-set of the real relevant documents set due to the low search engines database overlap. Meta-search engines which query many search tools at the same time can improve the completeness of the results.

3.1.2 Social Approaches
First of all, many search engines do not only use a classic statistical analysis of documents contents to index but also introduce a popularity of documents measures (either considering the number of incoming links or the visit frequency) like DirectHit (http://www.directhit.com). A more recent approach, HumanLinks system (http://www.human-links.com/), uses the Peer-to-Peer (P2P) protocol to allow a user to search in documents which have been shared by other Net surfers. Concerning the search result management and visualization, the VR-Vibe (Benford, 1995) system can be underlined. Besides this information retrieval task, a browsing task is implied in the search task.

3.2 Helping users to browse information

3.2.1 General Approaches
Browsing can be described as following available anchors (or links) from a source document to access a destination document. This succession of visited documents implies that the main issue is the cognitive effort it begets. Indeed, users have to remember the history of visited documents to avoid re-visiting it or getting lost in the hypertext. To limit this cognitive effort, browsers have introduced a history manager that presents all the latest visited documents. These histories are commonly represented through a (sorted) list. More visual approaches use thumbnails to represent each visited document (Hascoët, 2000). Furthermore, to avoid the user getting lost in the hypertext, some approaches propose a visualization of the navigation with its local hypertext through document that are linked to each visited document (Hendley, 1995). Approaches which recommend the user links he should visit such as WebWatcher (Armstrong, 1995) can also been quoted.

3.2.2 Social Approaches
Some projects have been developed such as annotation systems like Pharos (Bouthors, 1999) which allow a user to get information from others. Some other systems use the previous navigations to orient the current user through anchors during his browsing task like Broadway (Jaczynski, 1997), Letizia (Lieberman, 1997). These approaches only help a Net surfer punctually during his browsing task. Helping users should be done in a deeper way. Indeed, to really help a Net surfer to find relevant information in the long term, we assume that such systems must also make his domain knowledge evolve. Thanks to a better domain knowledge, the Net surfer should query search engines in a better way that limits some precedent issues. Moreover, due to the poor social characteristics of the Web, Net surfers cannot profit from other users’ knowledge and search
competence. Finally, developed systems only take into account one of the information seeking aspects (information retrieval, browsing or visualization of search results). Unfortunately, the user’s information need is still the same during all the seeking task. So, using independent tools implies that each system models the user in a different way because they only have a part of the information characterizing the user. So, an integrated system, which helps the user during the whole information seeking process, could be a more suitable solution than independent systems. As a solution, this paper presents a social and an adaptive system that helps the user browsing the web task and making his domain knowledge evolve in the long term.

4. A SOCIAL AND AN ADAPTIVE APPROACH

This paper proposes an approach which aimed at helping a Net surfer during all his information seeking task on the Web. It is based on a social system which automatically shares information between users. Three features which are naturally integrated into the search process without changing a lot the user’s habits are proposed:

- A visualization interface that gives the Net surfer a synthetic view of his search result,
- A recommendation system that gives new documents to the Net surfer according to his point of interest. It will help the user in the long term to make his domain knowledge evolving,
- A second recommendation system that gives new trails to relevant documents during his browsing task.

These features concern the information seeking process as a whole. The visualization interface concerns the information retrieval phase. We assume that the way to help a Net surfer querying a search engine is to help him to understand the search result rather than featuring a new search engine. Indeed, we assume that no search engine will cover the entire Web. The first recommendation system concerns the long term knowledge. Indeed, we assume that the privileged way to make a user more efficient in seeking information concerning his points of interest is to make his domain knowledge evolve. Thanks to the recommended documents, the Net surfer may construct his own tacit knowledge through new information. At last, the second recommendation system is aimed at helping the user to find relevant information during browsing. The recommended documents are not necessarily connected to visited documents like general approaches (Letizia, Broadway...). The two latter features are based on a social system. All the information recommended to the user is coming from other users. So, an application of such an approach can be seen in organizations which want to optimize the Knowledge Management through an information spreading process. The proposed features are proposed in an integrated system. This aimed at avoiding multiple representations of the Net surfer’s characteristics which can be shared between features.

4.1 A Visualization Interface to Manage Search Results

Querying a search engine may generate a high number of documents that can be hardly manageable for a user. To palliate this, a visual information retrieval interface (VIRI) is supplied. A visual information retrieval interface is aimed at presenting the search results as a whole and at managing them by spotting the relations between documents. This visualization is based on the fact that the Net surfer commonly uses few keywords in his query (Spink, 2002) and that he has to understand relations between retrieved documents. For this, the proposed interface combines colors and a specific spatial layout (Figure. 1) (Chevalier, 2000). It is based on the HSV color space model. Through this VIRI the relations between the documents and the query keywords are visualized. As a query is composed of about 3 keywords, three axes are used to visualize information; each one of them being assigned to a specific color (red, green or blue). Each plot of the interface corresponds to a specific combination of the different axes and so to many retrieved documents. The color of a plot results from the mixing of each color corresponding to each axe. The value (or intensity) of the plot’s color corresponding to any axis is relative to the importance of the query keywords assigned to this axis. If the query has more than 3 keywords, the Net surfer can combine many keywords, via boolean operators, on the same axis.

A qualitative and a quantitative study made on 12 persons shows that the combination of visualization criteria (layout and color) helps users to evaluate the relevance of each query terms in retrieved documents. A minimal experience is also required to manipulate the interface in an optimal way (3D visualization).
4.2 Recommending Information to the Net Surfer

The two recommendation approaches are based on the Net surfer personal information space (PIS). Indeed, all Net surfers save, bookmark or organize relevant information in their folders on their hard disk or their bookmark hierarchy for instance. Concerning the information seeking process, the proposed approach is based on the bookmark hierarchy as PIS (Abrams, 1998). Note that the supplied features to the user could be adapted or combined with any other PIS.

However the recommendation approaches are based on implicit information rather than explicit one because it does not require more cognitive effort for the user to supply it (Alton-Scheidl, 1999). Furthermore, the interest of using bookmarks is that through such an organization the user constructs a representation of his own points of interest. It is also important to underline that the insertion of a bookmark into a specific folder of the bookmark hierarchy (BH) results from a cognitive effort (Rücker, 1997). The latter can also imply a relationship between the document and the topic related to the folder. Thanks to the Net surfers’ PIS two recommendation features have been proposed.

4.2.1 Recommending Documents According to the Net Surfer’s Information Points of Interest

The first recommendation feature concerns directly the Net surfer’s points of interest (POI). Indeed, we assume that to help the user to maintain or to construct his own knowledge related to his information needs in the long term is to give him new relevant information regularly. To do these recommendations, the proposed approach exploits a representation of the users’ POI through a ‘profile’. A profile is a characterization of the Net surfer’s PIS. The solution proposed here is to recommend information to the user according to his PIS thanks to his profile according to a social process. The latter is based on the use of all documents that have been visited by any user as inputs. Using all the visited documents allows the recommendation process to profit from the search competence of each user. At the same time the system can have access to documents that are located in the Hidden Web. Furthermore, since the user only bookmarks or saves less than 50% of the relevant visited documents (Rücker, 1997). The use of all the visited documents is adequate.

To recommend documents, the process has to match each Net surfers’ POI (represented by a profile) and visited document to detect those that can be relevant. The way the profile is constructed depends on the way the PIS is organized. Most of the time, the latter is organized as a folder hierarchy. Each folder corresponds to a specific users’ POI. The proposed profile representation is naturally taking into account this organization. Each folder of the PIS is, so, characterized in a singular representation that reflects the content of documents it contains. To allow the user to access this representation (to construct a query related to one of his POIs for instance), it has to be understandable and readable. In the proposed approach, we have privileged the comprehension of the user. The Rocchio 16/4 representation (Rocchio, 1971) has been selected rather than the K-Nearest-Neighbor, Megadocument, SVM…, even if it is not the most efficient one (Sebastiani, 2002). The Rocchio representation consists in a term weighted vector. For a specific folder of the PIS, the Rocchio vector contains weighted terms that depend on:

- its positive importance in documents in the folder and in the (recursively) sub-folders to take into account the specialization/generalization organization of folders in the PIS,
- its negative importance in documents not located in the corresponding folder and its sub-folders.

The importance of a term in a document will commonly depend on its discriminant power that can be computed via the standard tf.idf measure (Baeza-Yates, 1999). To make the Rocchio vector optimized, it should only contain terms really representative of a folder content that is to say with a high $\chi^2$ value for instance. The profile is thus composed of a hierarchical organization of Rocchio vectors, each vector...
corresponding to a specific folder of the PIS. At the same time, each Rocchio vector is associated to a threshold that indicates the minimum similarity needed to decide whether a document is relevant to the corresponding POI. This threshold can be computed regularly without high resources requirement (Chevalier, 2003). At the opposed side, each visited document is indexed in the common way (Baeza-Yates, 1999). To detect if a document is relevant to a specific user’s POI, a matching phase has to be applied. To this, a hierarchical matching method is realized. The matching process starts at the root of the users’ profile. A matching between the document representation and each Rocchio representation (related to each folder located at the root) is computed. If the similarity resulting from this matching is higher than the threshold associated to the Rocchio representation, the document is considered as relevant and the matching process continues to the sub-folders recursively. At the end, a document will be considered as recommendable to the folders that are the deepest folders in a path for which the document is relevant.

An evaluation made on a part of the OHSUMED document collection (4976 documents) shows that considering the PIS folders in a hierarchical organization and using a hierarchical matching process makes the system performances increasing: +43% for the macroF1 measure (0.752) and -68% for the time required to recommend 1639 documents.

Thanks to this recommendation feature, any user gains documents directly in his PIS. These documents are directly inserted into the folders for which documents are relevant. It allows the user to understand why the document is recommended to him. A feedback procedure is also proposed, that is, according to the Net surfer’s needs (insertion of a new bookmark, positive or negative judgment concerning recommended documents…) makes the Rocchio vector and the threshold correspond to a folder of his PIS modified and adapted. This feedback process makes the system adaptive in the way that the user’s profile is corresponding to his “explicit” POIs. So, the more the system is used, the more the recommended documents are relevant to the Net surfer.

### 4.2.3 Recommending Alternative Relevant Trails While Browsing

When a user is browsing the Web, he is looking for relevant information via following a link (also called anchor) from a source document to visit a destination document and so on. But through browsing, the Net surfer’s information needs are not explicitly formulated and he is generally alone facing the documents he is visiting. A recommendation feature that propose the user some alternative trails to relevant information according to his information needs deduced from his navigation path is presented. It relies on a social approach to gain exploit the practice and the experience of other Net surfers. To do this, the proposed feature exploits the knowledge directly deduced from the users’ PIS construction. Indeed, the relation between documents in a same path or more precisely in the same folder can be used to deduce similarities between documents. The interest of such a method is that the recommended relevant documents are easy to find and are not necessarily linked (via anchors) to visited documents as in common approaches like Broadway.

To identify the user’s information needs the approach is based on a similar approach as (Lieberman, 1997) in which the Net surfer’s information needs are deduced from visited documents. The proposed approach relies on a navigation profile which is evolving during the navigation. The navigation profile contains an ordered set of visited documents and a set of potential recommendations related to the navigation.

To recommend documents that are relevant to the current navigation, the organization of all the users’ PIS is observed. To obtain a synthetic representation of all the available PIS, a multitree representation (Furnas, 1994) is used. A multitree (Figure. 2.) is a representation in which the objects (all the documents contained in all the PIS) are linked to all the tree nodes (folders of all the PIS). Thanks to this multitree some relations between documents can be exploited. To do this, the system uses the notions of “filiated” document: a “filiated” document of a document d is a document that appears in the same (sub)trees as d. Thus, in figure. 2., the filiated documents of document D4 are {D1, D2, D3} due to the trees called “Java programming” and “Programming language”. Thanks to these filiated documents, a conceptual distance is computed depending on (Chevalier, 2003):

- the average distance between a filiated document and the visited document. In figure. 2., the document D3 (distance = 0.5 folders [(0+1)/2]) is closer to D4 than document D1 (distance = 3 folders).
- the proportion of the users having, in their PIS, the visited document in a same tree as the filiated document. In figure. 2., the document D3 (2/2 users) have a higher proportion of the document D2 (1/2 users).

The main limit of such an approach is that it can only be applied to one visited document. So, to make the recommendations relevant to the Net surfer’s navigation, the proposed approach uses a set of recommendable
documents. This set of potential recommendations is composed of \((document, weight)\) couples in which \(document\) represents an identified document that is potentially relevant to the user’s information needs and the \(weight\) corresponds to the (system) matching value between the document and the navigation history as a whole. Each time a user is visiting a document \(d\), a set of weighted documents that are relevant to \(d\) is computed, which is used to make the set of potential recommendations evolve.

![Root-PIS1](Java programming UML C++ Java Examples) Root-PIS2

Figure 2. A multitree representing two Personal Information Spaces

According to this navigational profile, the highest weighted documents from the set of potential recommendations can be regularly proposed to the Net surfer. This approach makes documents that are relevant to a high number of visited documents more important for the user’s navigation.

An evaluation made on 21 PIS (5098 bookmarks = 4566 unique document) shows that more than 80% of the recommended documents are relevant to a single visited document. Furthermore, after 4 visited documents, most relevant documents are situated at the top of the recommendation list.

### 4.3 Implementation: the Easy-DOR System

Easy-DOR, standing for “Easy DOcument Retrieval”, is an integrated system that assists the user during his information seeking process. It is based on a proxy architecture that logs all the visited documents. To obtain the users’ bookmark hierarchies, the system proposes a bookmark server that permits the user to access his bookmark hierarchy wherever he connects. The visualization interface is applied to an integrated search engine that allows users to query the bookmark database. The recommendation feature related to the domain knowledge indicates relevant documents directly in the user’s bookmark hierarchy folders. The recommendations related to the navigation are presented in an independent window to avoid disturbing the Net surfer during his navigation. The prototype has been implemented in Java and is a stand-alone application. Some screenshots are available at [http://www.irit.fr/~Max.Chevalier/easydor.htm](http://www.irit.fr/~Max.Chevalier/easydor.htm).

### 5. CONCLUSION

Thanks to all of these features, the user can be assisted during his information seeking task as a whole in order to make it more effective. But searching information on the Web is so important that helping the user could only be envisaged through a global approach. Indeed, solving any issues like information retrieval without taking into account the other ones (like navigation) is not effective to help a user in the long term. We assume that an integrated system like the proposed approach may be an effective complement to most of independent tools.

The proposed approach is based on the idea that the user must be reinserted into his social context. It takes into account some of the issues he can encounter when he search information on the Web, from the lack of knowledge to the management of search results. It exploits users’ information retrieval skills and knowledge to help a third part. Furthermore, the proposed adaptation allow the system to follow the evolution of the user’s information needs.

The Easy-DOR prototype underlines the feasibility of such an approach and the experiments done confirm the interest of the proposed features. Nevertheless, some evolutions could be envisaged. First of all, the proposed approach should integrate a visualization of the navigation history to help the user while he is browsing. Concerning the search results visualization, the system should provide many other visualization...
metaphors to facilitate all the users’ tasks according to their skills. Moreover, the visualization is limited to the integrated search engine only. In the long term, assuming the user queries specific search engines, the visualization has to evolve towards a meta-visualization interface. Concerning the recommendation features, some further experiments have to be done to ameliorate the hierarchical representation of PIS. Furthermore, more indicators should be introduced to improve the user’s information needs characterization during the browsing task. More evolutions of such an approach can also be seen in People-to-People paradigm for instance.

REFERENCES


