IADIS International Journal on Computer Science and Information Systems
Vol. 6, No. 2, pp. 46-68
ISSN: 1646-3692

EXTRACTION OF CONTEXTUAL ASSOCIATIONS TO SUPPORT USER IN A TASK OF INFORMATION RETRIEVAL BY NAVIGATION

Jean Caussanel and Ali Mroué
Laboratory of Sciences of Information and of Systems-LSIS UMR 6168, University of Paul Cezanne, Aix-Marseille III, Av. Escadrille de Normandie Niemen 13397

ABSTRACT

Information Systems adaptation to its users requires to be able to identify real uses made of the system. The behaviour models, extracted from System-User interaction traces, provide representations of the real tasks performed in the Information System. We propose to exploit these representations in order to extract knowledge that can, in turn, help to support users in their activities or to adapt the system to emerging activities. This paper describes such an approach we applied in the field of Information Retrieval by navigation in hypermedia. The behaviour models consist of navigational patterns and extracted knowledge is a thesaurus whose content is strongly related to the context of the available information. The paper is focused on the process of relation extraction between terms. It also presents experiments carried out for testing relevance of these representations. We show that by using the associations of the thesaurus and keywords describing the user's needs, it is possible to predict the requested page from the first pages of navigation. The results are evaluated by measuring the predictive capacity of such relationships in a perspective of pages recommendation.

KEYWORDS

User behaviour Modelling and Simulation, Navigational Patterns, Information Retrieval in Hypermedia, Contextual Knowledge Extraction.

1. INTRODUCTION

In order to adapt an Information System (IS) to its users, it is imperative to identify existing usages of this IS. According to these different usages, it's possible to design, or re-design, the IS so that it offers an adapted structure to its user profiles. These profiles can be captured by the way of users' behaviour models acting in the IS. That also gives the opportunity to use these models for simulating behaviour of categories of users facing with changes in the information organization. Therefore, modelling and simulation of users' behaviour appear as

an important issue in the domain of adaptability. In current computing environments, each user action, each event is recorded in a log file. The traces are valuable because they provide information about learning pathways, actual usages, emerging usages, often different from those initially planned by the designers. The acquisition and analysis of users' trails can, in turn, allow to improve the information structure or processes of progress in applications. The issue that we address more precisely in this paper is related to the interpretation of traces in order to analyse, and to model, the users' behaviour. We argue that it is possible to extract from the System-User interaction traces some knowledge involved in user decision. In our case, this knowledge is extracted as a set of associations between terms and built from the more significant behaviours. Our hypotheses and methods are experimented in the field of Information Retrieval (IR) in hypermedia. However, it is important to keep in mind that we designed our approach for a broader framework. We address all the tasks whose cognitive models are structured as a tree in which the user navigates with a decision process based on his/her knowledge and system status. Due to this aim, we rejected approaches only focused on the task of Information Retrieval (IR) which are difficult to generalize.

The remainder of the paper is organised as follows. In Section 2 we give and justify our choices for an approach of *"user adaptation"*. Then, we focus on the primary subject of the work presented here: the acquisition and modelling of knowledge related to the task of Information Retrieval by Navigation in Hypermedia (IRNH[1]). We present related approaches and give for each of them, basic differences with our position. Section 3 provides a description of our method for contextual association extraction on the basis of interaction traces. The implementation of the method and generated models are presented on a concrete example. In Section 4, we evaluate the relevance of extracted associations by using them for piloting a virtual user and for predicting the target page of users on the basis of the very first visited pages. The results are compared with the real choices we have in the log files.

Finally, we conclude and give our next goals to further improve the knowledge extracted from the task but also user support.

2. POSITION AND RELATED APPROACHES

2.1 User Simulation in IRNH[1] Task

Although this is a very common task, the task of Information Retrieval by Navigation through a Hypermedia has been the subject of a relatively small number of works in the field of modelling and simulation. This observation concerns the cognitive and computable models, those which offer a comprehensive and autonomous behaviour model. There are many studies that have sought factors that influence the decision in the task of IR [Herder 2006] but very few offer a **cognitive and computable decision models**. Such a statement, already made several years ago by [Tricot & Nanard 1998] remains relevant. The only two approaches which fall in this category are SNIF-ACT [Fu & Pirolli 2007], CoLiDeS and CoLiDeS+ [Kitajima et al. 2000] [Juvina 2006]. The studies such as those conducted by Herder [Herder 2006] aim at bring out the factors influencing the cognitive processes but do not provide a

[1] IRNH: Information Retrieval by Navigation in Hypermedia

computable model. When Juvinas sought to implement the cognitive principles that he has highlighted, he has done by proposing an extension to CoLiDeS namely CoLiDeS+ [Van Oostendorp & Juvina 2007].

One of the reasons that can explain such a situation (little number of complete models) is related to the complexity of cognitive processes involved in this task. Ideally, a good model should reproduce step-by-step process by which information presented on the screen is received and analysed in order to decide the next action to achieve. This requires not only a user model (with all kinds of knowledge, memories and decision processes) but also a representation of the system and its environment. For example, CoLiDeS [Kitajima et al. 2000] is based on a model of text comprehension and SNIF-ACT relies on a comprehensive architecture of cognitive modelling and simulation ACT-R [Anderson et al. 2004].

In the field of Information Retrieval [Santos & Nguyen 2009] proposed their own user model: the IPC (Interests, Preferences, Context) User Model. However, it is not a cognitive model of the task with a representation of the knowledge and processes mentioned above. The IPC model associated with each user, addresses three types of knowledge: *(i) Interests* as for direction of individual's attention; *(ii) Preferences* as for possible actions to carry out; *(iii) Context* for user's motivations behind a specific goal. The only action available to the model is the submission of a *query* and two derived actions namely *filter* or *expand* on query results. The model is effective because it is dedicated to a task using only a search engine. The needs are not the same as for a cognitive model of the task of Information Retrieval by Navigation.

In a cognitive perspective, we can schematically represent knowledge involved in the task of IRNH[1] as shown in Figure 1.

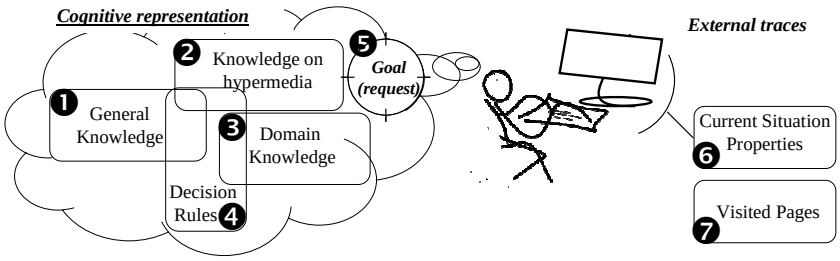


Figure 1. Knowledge and information to be tackled for IRNH task modelling

User's knowledge (on the left) can be separated from visible information about its activity (on the right):

1. **General knowledge** which deals with knowledge relevant for any situation of life (it refers to the knowledge of *semantic memory*). For example, the fact that *writer* and *author* are two related notions;
2. **Knowledge on the hypermedia** which deal with his/her experience with this hypermedia or another one on the same topic. For example : *information is organized by authors, or by type of books (novels, biography, non-fiction,…), topics, ...*;
3. **Domain knowledge** which deal with his/her general knowledge on the field. For example, the fact that *Stendhal is a French writer of the 19th century* ;
4. **Decision rules** which deals with rules used to make a choice between available actions being given current knowledge and current situation;

5. **Goal request** which deals with his/her request and its current status of progress e.g. "*now I look for the last volume of the Italian Chronicles*" (I already found the other)
6. **Current situation properties** which deal with the various choices which being presented to the user at a given step, e.g. *link towards the author, search engine, the authors of 19th…*
7. **Visited pages** and all logs about his/her navigation in the hypermedia i.e. page he/her visited prior to the current one.

While some of these elements can be easily acquired, such as the previous *visited pages* and the *current situation* (6 & 7 in the figure 1) or, under certain conditions, the keywords of the *request* (5), the others are hard to collect due to the broad scope of the task of Information Retrieval. For example, the skills of text comprehension for every field of application, and the corresponding knowledge of the semantic memory, are one of the main requirements for such a model of task. Therefore, state of the art about user simulation for this task shows:

- on the one hand, some works that offer a general model of the task, independent from the different fields of application, but designed at a high level of abstraction. In these cases, no implementation is provided in a computable form;
- on the other hand, there are some works that chose to limit the representation of knowledge to relatively simple forms like associations between terms that will be used to select the next action to undertake.

The EST (Evaluation, Selection, and Treatment) model is an example of the first category (Figure 2) [Dinet & Tricot 2008]. Proposed in the domain of Cognitive Psychology, it offers information about the processes involved in the task but it is abstract, with no implementation. Authors designed the model as a sequence of three phases of skills: *Evaluate* (relevance of information); *Select* (information in the document); *Treatment* (of selected information). However, the aim of the authors was to provide an explanatory conceptual model, not a simulation model. Therefore for someone who would aim develop a computable model from the EST model, it would be necessary to define:

- how one can identify and represent the goal of the research ;
- how one can represent the current situation ;
- how one can identify and represent the research plan ;
- how one can identify and represent the available knowledge.

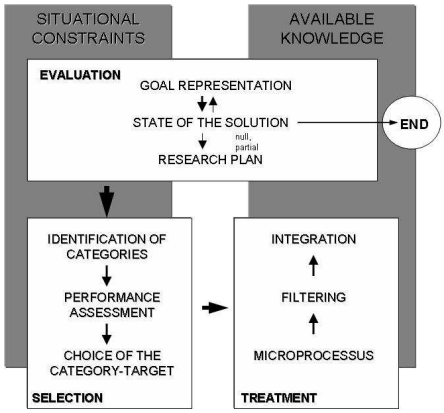


Figure 2. The phases of the EST model

On the contrary, approaches such as SNIF-ACT and CoLiDeS in the field of Artificial Intelligence have a modelling approach oriented towards simulation. In this domain, consistency of models is evaluated through their capacity to mimic the human behaviour in a same situation.

In order to be able to reach this goal (simulation), the decision modules compare attributes of the situation with a representation of goal using some pre-stored knowledge (general and domain knowledge). In the field of Information Retrieval in hypermedia, that usually leads to compare some terms: a set of selected terms in the current page and terms defining the goal (directly given by user or collected by another means).

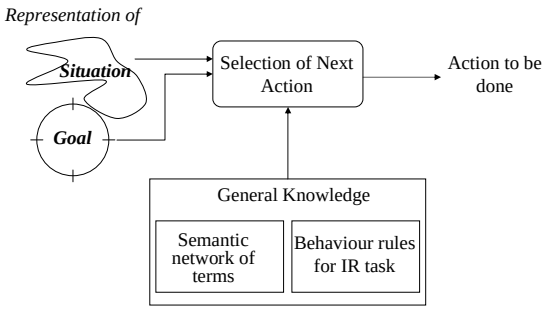


Figure 3. Generic principle of decision model in IR task modelling

Figure 3 above gives a synthetic schema of these models of decision for Information Retrieval (IR) task. This comparison is used to:

- determine the level of achievement of the goal ;
- decide of the next action to proceed. Depending on this evaluation, the user can :
  - *complete the task*, either because he failed or because he succeeded;
  - *continue the search* due to current results assessed as "not enough relevant" ;
  - *go back* to a previous page ;

The choice of the action to be done is based on the comparison of terms describing the current situation, some terms describing the goal and terms associated with a given action (a premise of a production rule for example). The quality and the realism of the model are strongly related with the quality and realism of this comparison. However, if the comparisons are limited to the keywords of the query and the terms appearing on the current page then the comparison will never be positive. Therefore, as the user would do, it is necessary to use other terms related to the first by semantic associations.

The SNIF-ACT model proposed by [Fu & Pirolli 2007] directly tackles this problem. It proposes an approach that is both realistic and effective while relying on a cognitive theory in Information Retrieval (the theory of Information Scent [Pirolli & Card 1995]). The relationships which are exploited are represented in a large spreading activation networks automatically constructed from on-line text corpora. This network provides a semantic distance between pairs of terms. In SNIF-ACT 2.0, corpora are samples of Web documents locally stored. If terms to compare are not present in this corpus they use a technique that queried the Web for statistics about words. With this general network, a priori constructed, SNIF-ACT can provide predictions of users' behaviours in various domains.

As for SNIF-ACT, CoLiDeS (CoLiDeS : COmprehension based Linked model of DEliberate Search)[Kitajima et al. 2000] is both a model and an environment of simulation for

the activity of a user's traffic within a hypermedia. The cognitive model used in CoLiDeS is based on an analogy between the processes of texts comprehension and those used in information retrieval on the web [Kitajima et al. 2000]. The search process is divided into a set of cyclical phases. In each of these phases, some comparisons are made on the basis of three factors, but we focus here on the first one, which is a measure of semantic distance. Unlike SNIF-ACT, CoLiDeS uses an algorithm of LSA (Latent Semantic Analysis, [Landauer & Dumais 1997]) to evaluate the semantic distance of words or goals and to choose the action to be fired. The LSA also uses a corpus of documents to calculate the properties of terms and evaluate the distance.

These networks or properties of terms (in CoLiDeS or SNIF-ACT) are considered as a kind of knowledge relevant for this kind of task. Depending on the corpora that they use for extracting associations, it may concern general knowledge enhanced with some domain knowledge. Such general corpora can't include highly specific information about a given organisation. However, each IS has its own vocabulary related with very contextual information of a company, a department, a group or a community. Our aim is to address this *highly contextual* information like the fact that *M.X* is an associate of *M.Y* or that *Z* is a researcher of the team *T-Team*. We argue that it is an essential part of knowledge for simulating a user seeking information in a specific context. These contextual information or knowledge have already been addressed in the field of Information Retrieval. We will now present these works and our position with regard to those which could be exploited for our own purposes.

2.2 Contextual Knowledge for the Task of Information Retrieval

Our proposal is to enrich existing representation of knowledge in computable models of IRNH[1] task by the means of knowledge extracted from traces of activities collected in the IS (Information System). Knowledge addressed in this case is deeply related to a specific organisation, company or other specific context. Since the representations we use are based on a semantic network, we can consider this network as a contextual ontology [Bouquet et al. 2003] [Benslimane et al. 2006]. Due to the importance of context in the process of selection of information by the user, researchers have sought to represent the context and to use it in the Information Retrieval System IRS[2] as an ontology [Tamine et al 2010]. Although we are interested in relevant associations in a given context, our position is distinguished by the fact that we consider the information in context and not the context itself. Context is a notion variable in nature, with multiple levels, multiple aspects, and no single definition as noted by [Cool & Spink 2002] or [Polaillon et al. 2007]. In [Tamine-Lechani et al. 2010], the authors proposed a comprehensive taxonomy of context aspects divided into five main categories (*Device Context*, *Spatio-Temporal context*, *User Context*, *Task Context*, *Document Context*) and some subcategories like *personal* vs. *social* context for the user context.

Nevertheless, representations of the context that can be found in the IRS[2] often cover more than one of these categories. For example [Hernandez et al. 2007] address two levels of context: a thematic profile of the user (*User Context*) and a set of information related to the request (*Task Context*). These same two dimensions also exist in the works of [Mylonas et al. 2008] since the context, seen as a set of semantic relationships, is used to represented

[2] IRS : Information Retrieval System

preferences of the user and to modify semantic relationships exploited at research time (*global context* or *instantaneous context* for [Polaillon et al. 2007]). But the authors take a much more imprecise definition of the context: *"the interrelated conditions in which something exists or occurs"*. They used fuzzy set to determine the extent to which components of domain ontology will be considered as relevant, or not, in order to expand or reduce the set of concepts related to the research. [Campbell et al. 2007] and [Liu & Chu 2007] focus on the context related to the task, the *Task Context*, according to the categories of [Tamine-Lechani et al. 2007]. For example, [Liu & Chu 2007] determine, among a set of pre-established scenario of research, the one to which the current research may be assigned to. Finding such an associated scenario would enable the system to enrich the query with some related terms of the scenario. [Chaker et al. 2010] go beyond the scenario by proposing an architecture of contextual IRS^2 based on a more general concept, the *situation*, incorporating a representation of user task and the environment. This approach meets our objectives since it is not only dedicated to the framework of information retrieval. However, at this stage, the creation of the set of reference situations requires a priori definition of rules and features (the components of the *situation*) whereas we aim to elaborate models based on users' real practices.

We differ from these approaches in the way we address the context. We see the context as a set of information which implies a particular semantic for a set of variable. For example, let's consider the concept of *department* (the variable). The sense of this concept (the semantic), will change depending on whether it is in the context of a *company* or *university* or an *administrative division* of a state (the context). Consequently we do not pretend to model the context itself (what is a *company*, a *university*, …) but the information in context : a *department* of a particular company and terms which are related to "department" in this particular company. An association between terms may be meaningful in a given company and meaningless in another. It is closed to what [Bouquet et al. 2003] denote as *local models* which provide a local view without compliance with models dealing the same concepts, in another context: "*We say that an ontology is contextualized, or that it is a contextual ontology, when its contents are kept local (and therefore not shared with other ontologies)…*".

Ontologies developed to support the task of Information Retrieval, are designed to be as general as possible in order to be reusable in different fields. Thus, it's not necessary to rebuild some "costly" knowledge structures for each particular case in which IRS^2 is deployed. However we believe that some requests cannot be satisfied if the IRS^2 has no representation of very specific information about the organization it serves. Unfortunately, unlike general knowledge, contextual knowledge has to be acquired in each case. Therefore, to make that possible (and realistic), the process of extraction has to be fully automatic.

Our approach aims at extracting this set of knowledge, as a thesaurus, less rich than ontology, but automatically built and highlighting some associations between terms strongly related to the actual tasks performed by users of the system. The terms that are linked appear in the same class of navigations (represented by a navigational pattern), which should relate to an existing task in the system. The association between terms is created because of an existing navigational pattern and only on this basis. In other words, the relationships are not based on some theoretical knowledge from general corpora but collected from users' existing practices.

Other differences lie in the nature of knowledge models and the context representations that we used, and about the way we elaborate them. In [Hernandez et al. 2007] the model of context and knowledge are based on ontologies, previously developed. They are produced in a semi-automatic manner from corpora of documents on a theme and sometimes also from synthesis of existing ontologies. Even *lightweight*, these structures are complex to elaborate

and maintain [Bouquet et al. 2003][Sondhi & Chandrasekar 2010][Santos & Nguyen 2009], requiring a very time consuming effort (about 30 hours for the case studied where a corpus of relevant documents was already available in [Hernandez et al. 2007]). Moreover, implementation of this approach in domains where number of document in corpora are relatively low (website), and the performed tasks are poorly identified, may lead to some results much less satisfactory than those mentioned in the article (too few occurrences of concepts and terms). The same observation could be done for other approaches such as [Eirinaki et al. 2004] which exploit (and not extract or build) some existing ontologies to enrich models of navigation and providing, by this way, a support with recommendations to the user. Our approach is different. The associations between terms are not based on an existing ontology of domain or again on some general or theoretical knowledge. They are based on real cases of navigation between pages: a class of navigations which is supposed to be the trace of a given task.

The method implemented by [Mylonas et al. 2008] also operates ontologies as models of domain knowledge with fuzzy sets as representations of context. In this case, thanks to a "*long term monitoring*" of user, the representations of context are produced by exploiting history of user interactions. But such approach which is basically "single-user" does not take advantage of all users experiences. In addition there is a risk that the user considers the system too intrusive and doesn't use it. It is the same for approaches which base the provided knowledge (contextual or not) on tags added to documents viewed by the user communities [Conde et al. 2010] [Xu et al. 2008]. This can work on sites dedicated to this type of exchanges but can fail on the IS of a company due to the nature of relations between users.

The works of [Campbell et al. 2007] and [Liu & Chu 2007] are closer to our approach since they exploit historical data to extract associations. However some significant differences exist. In [Campbell et al. 2007] the process of model building is not fully automatic. Some information has to be added or validated by the user himself. We attempted to avoid such request to the user which he could see as too much intrusive. In addition, the relationships addressed by Campbell et al. deal with documents and not directly with the terms involved in the task. In both approaches, the order of retrieved documents is not exploited, although it can provide information about the task. The authors also made the choice of a single-user context-based monitoring tools installed on the client side. The exploitation of the interactions of each user supposed that is possible to have means of ongoing monitoring on computer of the user. We have chosen a technique for data collecting and assistance from the server side. In addition to the fact that the client side is not always accessible (as for the web) it allows to prevent any intrusion on the client side and to take advantage of activities of all users.

Finally, in the approaches we just presented, the type of help provided relates most often to the requests submitted to a search engine whereas we aim at a navigation support. The contextual associations can also be acquired by exploiting the query logs [Baeza Yates & Tiberi 2007][Sondhi & Chandrasekar 2010]. The keywords that appear in a given query are considered as semantically linked. In this way, it's possible to collect a set of associations to form a semantic network of terms. But as we already mentioned, our approach is not only oriented toward the IR task. We are interested in assistance to all computer-mediated tasks. In computer science (and more generally), a task is most often designed as a set of steps structured as a tree [Hernandez et al. 2007][Chaker et al. 2010]. Support to user is then to assist him in the choice he faces at each node of the tree on the choice of action to be performed. The task of IRNH[1] meets these characteristics in contrast to IR processes which

exploit a search engine (see for example the overall architecture of IR application in [Santos & Nguyen 2009]).

In presenting these related works we wish to show how our approach is different, and probably complementary, from existing approaches. Our particular position can be summarized in a set of constraints to be fulfilled simultaneously in order to meet our goal. These constraints deal with the nature of information collected and how to acquire it. We focused on a particular type of information:

- highly contextual to the Information System or to the organization ;
- only extracted from the existing practices and common to all users;
- acquired and modelled automatically, without involving users at any stage of the process ;
- coming from, and dedicated to, users' navigation (not for search engine).

These objectives and these constraints being characterized, we will now describe our method to extract contextual associations from traces.

3. CONTEXTUAL ASSOCIATIONS MINING FROM TRACES

For the one who seeks to offer model close to real behaviour of users, it is necessary to take into account real uses observed among users. Traces of user-system interactions are one of the best ways to transparently acquire information on actual behaviours of users. Web environments provide easy access to these traces of interaction but they yield a simple history of accessed pages. However, it seemed to us interesting to make the maximize use of this source of information on real practices. Based on these raw data, we derived some patterns of navigation from which, now, we look for some contextual knowledge. To get realistic behaviour, a model of user should exploit some semantic relations (between words) which are meaningful only for a very particular IS (a particular company for example). The task involved is the task of Information Retrieval by Navigation in Hypermedia.

In this section we briefly remind the principles of the discovery of navigational patterns from traces before describing *CATMiner* the method for extraction of contextual associations that can enhance cognitive models of this task.

3.1 Principles of *CATMiner*

The extraction of contextual associations is based on a set navigational patterns mined from log files of a given hypermedia. These patterns are identified using a clustering method (namely *PatMiner* for pattern miner) that we developed with the ultimate goal of improving cognitive models of IRNH[1] [Mroué & Caussanel 2009][Mroué & Caussanel 2010]. *Patminer* extracts classes of sessions and a representative session of the class which is the navigational pattern (we call it the prototype of behaviour). These classes of users' sessions are related to significant behaviours on the web site which should correspond to some tasks performed in the hypermedia. Our work addresses, above all, these tasks. In such a perspective, there is no difference between tasks consisting in searching for some publications of the researcher *"X"* and those dealing with articles of the researcher *"Y"*. In both cases the task is: *collect information about a researcher*. Therefore we chose to identify classes of pages on the web site based on their content or needs it fulfilled (figure 4). Our algorithms generally handle

theses classes of pages but it's always possible to access to the instance of pages in the class and vice versa.

For example we grouped all the personal pages of researchers within a class called "*Personal Pages*". When *Patminer* is analysing the log of the sessions it replaces each individual page by the generic class. In this way, two paths with a sufficient number of classes of pages in common may be considered similar, although in reality users have chosen different instances of pages (figure 4).

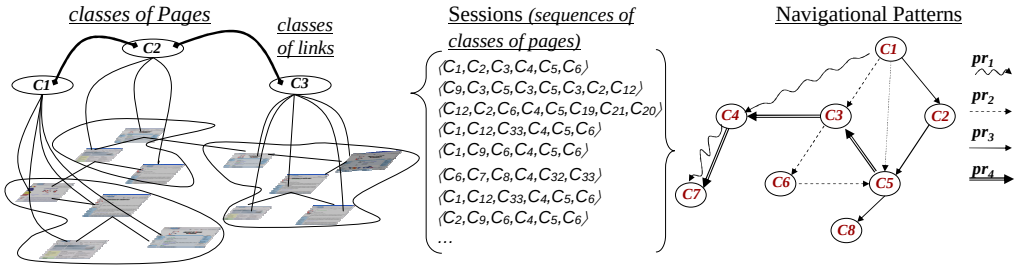


Figure 4. From normalized sessions to classes of behaviour (navigational patterns)

Each page is associated with one and only one of these classes. For instance, regarding the example that we exploit later, the web site of our laboratory, we got about 25 classes (against about 400 pages really accessible): *Laboratory Home Page*, *Personal Page*, *Publications*, *Research Reports*, *Teams*, *Members*, *Directory*, …

The number of resulting classes may seem relatively low. This comes from consistency of information available on the web site and the fact that some classes have a lot of members. For instance the class *Personal Page*, contains about 200 pages (one per researcher). The class *Teams* contains description of 7 teams with, each time, several pages of information. In addition to the classes of pages we used also classes of links. This is necessary when several pages of the same class C1 have links towards target pages belonging to a same class C2.

Let's remind that these prototypes are representatives of classes in which one can find the sessions either in their normalised version (the pages of a sequence are represented by the classes they belong) or in their specific version (the pages are those actually accessed). This means that search for associations can be performed on specific sequences and provide an association between terms of instances and in the same time provide an association between terms classes of pages or the classes of links.

The method for extracting binary associations between words is based on the two following assumptions (from now, we will use *CAT*, *Contextual Associations between Terms*, for the set extracted association and *CatMiner* for the algorithm which produce the *CAT*):

1. the words that are viewed by the user, and particularly the terms that are found on the links, have an impact on his choice ;
2. the words read by the user in its pathway within his session are related to his goal;

To be sure to use the terms actually read by the user, we have only considered - in the method described here - the terms that appear explicitly on the link. It is likely that the terms appear in close proximity also play significant roles.

To extract associations, the *CatMiner* uses, as input, the set of prototypes provided by the *PatMiner* clustering algorithm [Mroué & Caussanel 2010]. Let's call "Π" this set of pages.

The Figure 5 below, shows an example of Π, where the vertices are pages and edges are transitions between pages. The different kinds of vertices denote different prototypes.

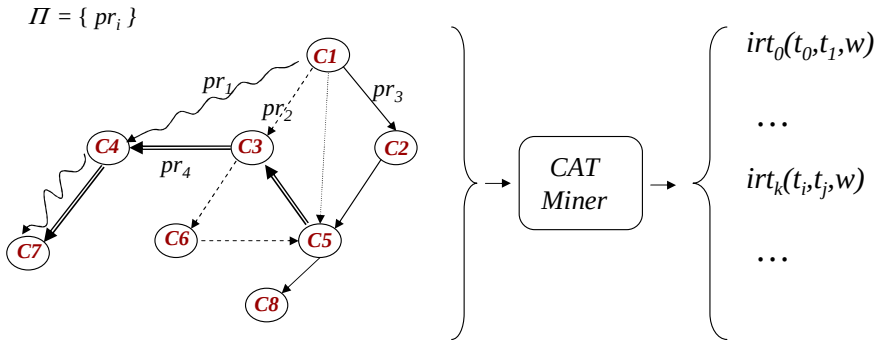


Figure 5. From navigational patterns to associations between terms

Each prototype pr_i of Π consists of n visited pages and a scalar value which represents the weight of this prototype in relation to all other prototypes:

$$pr_i = \langle C_1, C_2, C_3 \ldots C_n ; w \rangle$$

To go through the pages of this prototype, the user chose the links denoted by some terms in the web pages (we do not tackle case of pictures):

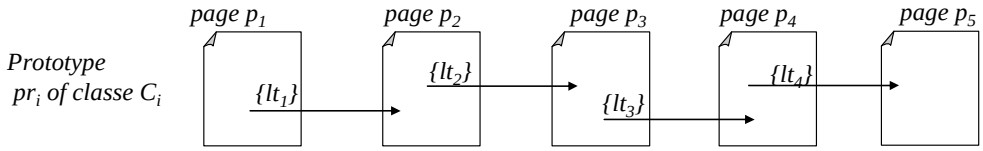


Figure 6. The terms of links between pages of a prototype are semantically related.

If we consider a prototype pr_i of n pages (as figured above) then the sequence of words selected by the user may be represented as follows:

$$Tpr_i = \langle lt_1, lt_2, lt_3, \ldots , lt_{n-1} \rangle$$

$$lt_i = \{t_0, t_1, \ldots t_k\},$$

where each $\{lt_i\}$ (*link terms*) represents the set of terms $\{t_0, t_1, \ldots t_k\}$ appearing on a link l_i of page p_i and leading to the page p_{i+1}. From this set, we build a set of binary associations between the terms of Tpr_i with $i \in [0;n-2]$. Semantic of the association is undetermined. Hence we take a general meaning for denoting it: *"is related to" (irt)*.

Each association is weighted according to the distance (the number of links to cross) between two terms t_i and t_j in the prototype pr_k. The own weight w_{pr} of the prototype is also taken into account. The more the prototype presents a high weight, the higher the weight of the relationship. The formula for the evaluating this weight between terms t_i and t_n in the prototype pr is given by the following expression:

$$w_{irt(ti,tn)} = w_{pr} - (w_{pr} * ((DecSim) * n-1)) ,$$

where w_{pr} is the weight of the prototype and *DecSim* (*0.1*) is the rate of decrease for similarity relevance applied for each (n-1) links existing between the terms. If the same association appears in another prototype the weights are cumulated.

We present the extraction of associations directly from the prototypes. But, in fact, this process is performed on the original sessions, on the basis of prototypes. This means that for each session of a class (and the associated prototype) only pages belonging to the prototype are considered.

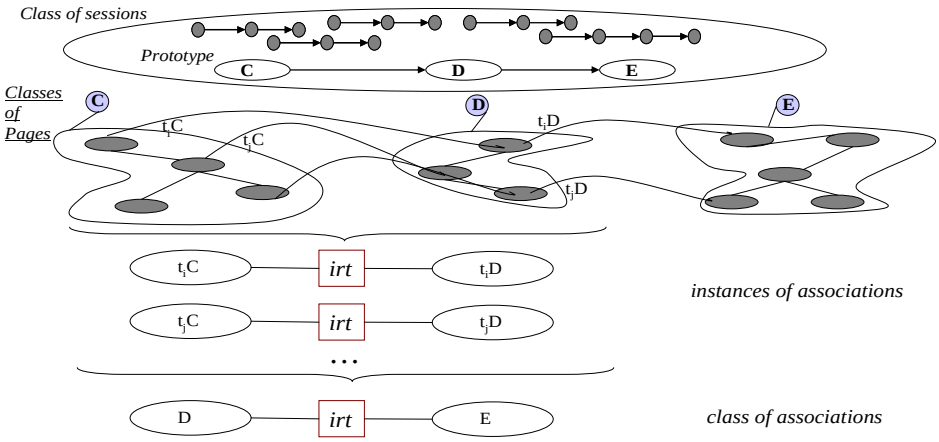


Figure 7. Class of associations

By using original sessions we are able to build associations between terms which are very specific to the site. Nevertheless, all of association instances coming from a same class of page are gathered in one (same) class of associations. The words we choose as an argument of the association are those assigned to the class of pages pointed by the links where the terms appear (figure 7). For example let's assume that *CatMiner* proposes the following relations:

irt(T_Team, Smith Personal Page),
irt(X_Team, James Personal Page),
etc.

We group these relations in a class whose the representative is:

irt(Team,Personal Page)

It is on the basis of this general relation that the weight of the association will be evaluated since it's based on the prototype which only consists of class of pages.

Finally, let's add that we also used the links on the pages, regardless of prototypes, for establishing more associations. We call these relationships, the structural associations since they are related to the structure of the web site. They are extracted on the basis of the existing links between the pages of the web site but we gave them a very low weight because we wish to emphasize, for this experiment, the part of decision related to the existing uses on the site rather than the structure of the site as it was originally designed.

3.2 Implementation

In the previous section we described the method for contextual knowledge acquisition and representation. We will now give an example of implementation of the method to show the

type of results produced. To conduct this experiment, we worked on the website of our laboratory. Besides the immediate availability of the information, this web site has other advantages for our experiment:

1. We are experts on its content and its structure which gives an advantage for behaviours analysis ;
2. It is a repository of information, which is limiting the types of tasks solely to the task of information retrieval;
3. It has no search engine which is leading users to navigate the site;
4. Finally, even if the reader does not know this particular site, he can easily imagine its structure and its content that should not be so far from the any other laboratory. This allows us to show anonymous information (out of respect from colleagues) without compromising the assessment of the results presented

By using the *CatMiner* algorithm on all pages of the laboratory and 110 prototypes, we obtained about 72 classes of associations between terms. These relationships form a graph in which the nodes, that is to say the terms, are connected by edges symbolizing an association between terms such as "*is_related_to*". It may represent a relationship of synonymy or hypernymy depending to the different cases. For now we can characterize it further.

The table below shows some classes of associations between terms with associated weight (rounded and normalized). The absolute value of weight is not very important because these values are mainly used to determine which association to select when several choices are possible.

First Term	Second term	Weight
Members	*Team*	*17.2*
Members	*Personal Page*	*15.0*
Directory	*Research Reports*	*0.12*
Directory	*Personal Page*	*10.7*
Articles	*Team*	*3.7*
Organisation	*Personal_Page*	*0.9*
Team	*Personal Page*	*17.5*
Team	*Member*	*6.5*
Master	*Lab's Life*	*0.1*
Organisation	*Research Reports*	*0.1*
Contacts	*Who's who*	*1*
Personal Page	*Directory*	*1.8*
Personal Page	*Team*	*10.0*
Conferences	*Members*	*1*
…		

Figure 8. An abstract of the association between terms of the site.

Let's notice that extracted associations are unidirectional due to the weight which may be different in one direction and another. Let's also remind that these associations are not between pages but between the terms. In the case of the site of our laboratory (as in many other sites) links are denoted by terms that are very close to the names given to classes of pages. It is usually a desired effect by the webmaster in order to give ideas on what can be found by choosing a given link (*information scent*).

Let’s consider association *irt(Directory,Research Reports)* or *irt(Personal Page,Teams).* Such a kind of relationship is related to the specific context of a Laboratory and to the specific context of THIS laboratory which chose to organize and linked information according to these terms. And yet, these associations appear to us as obvious. These kinds of relationships cannot

be offered by any existing ontology even though they are elaborated on the basis of domain oriented corpora. This is the same thing for a majority of associations extracted by CAT.

Extracted associations are represented and stored using the formalism of Topic Maps [ISO/IEC 1998], in the xtm format [XTM 2006]. Among the existing formalisms in the Semantic Web as RDF and OWL, Topic Maps is the one that seemed most appropriate because it gives importance to associations between topics as much as the definitions of the topic themselves. The Topics can be described by Uri towards pages of description, but it is not necessary to provide more information regarding their contents, properties structure, ... Precisely, we have no information about the terms (and related concepts) related through associations extracted.

The screenshot below shows the main part of relationships organized by a tool for Topic Maps management and visualization: Ontopia Vizigator (www.ontopia.net).

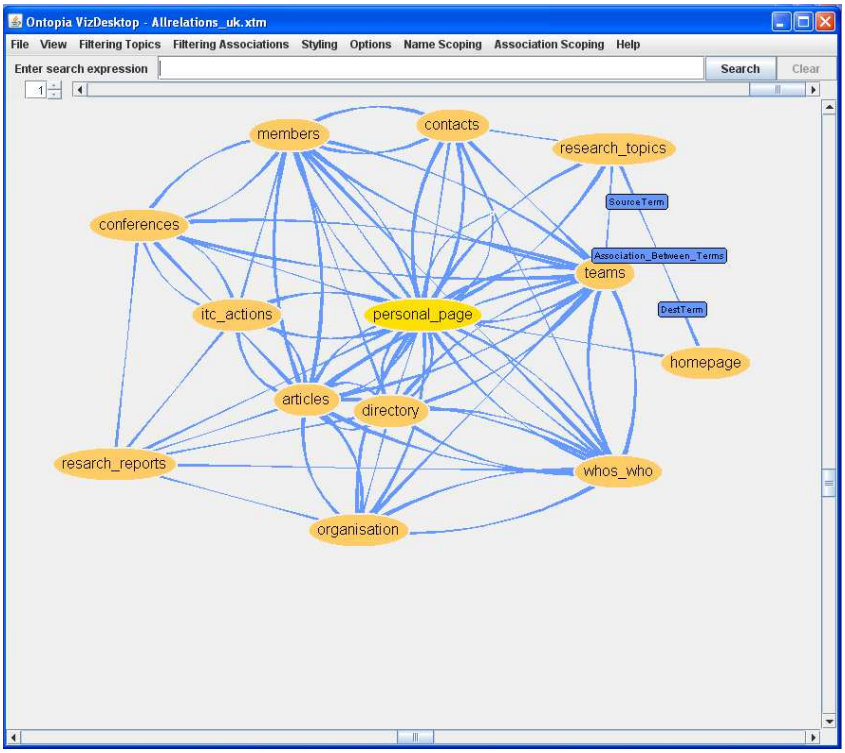


Figure 9. Topic maps of associations extracted by *CatMiner*

An example of association is unfolded between *Research_topics* and *homepage*. The advantage of using formalism like the Topic Map lies in the possibility of merging our network of associations with other representations (ontologies) described in compatible formats. Note also that it is always possible to export a Topic Map towards RDF or OWL representations. Moreover, this kind of formalism provides tools for researches and inferences that can be useful to provide recommendations.

We get, thanks to *CAT*, a set of associations between terms that come from actual uses of users and based on their experiences in the hypermedia considered. We will show in the following sections how this knowledge can be used in a goal to help the user in his/her task.

4. EVALUATION OF THE EXTRACTED ASSOCIATIONS

The previous sections have shown how to develop a representation of contextual knowledge on the basis of navigational patterns. However, usefulness of such representation has to be evaluated. Assessing the added value and quality of such information is not an easy thing to do. What is a good recommendation? What is the correct page to recommend and what is the best path towards a document?

To check interest of CAT relationships, we designed a first experiment (section 4.1) for a qualitative assessment: evaluation of the pathway taken by a browser robot, the Virtual User (VU), initialized with keywords describing the request. He navigates through the website, looking for a page having content related to the set of keywords;

The second experiment (section 4.2) deals with a test of prediction of a target page from the first pages of the sessions and on the basis of selected keywords. We sought to compare prediction based on the relationships with actual paths available in the log files. Since it is difficult to work with a great number of human subjects for this kind of experiments, we use existing logs on a Web site. Thus, we can provide statistical evaluation about prediction.

4.1 Navigation of a Virtual User

To understand our approach it should be noticed that this work falls in a larger framework of user behaviour modelling from interaction traces [Mroué & Caussanel 2009] and not only for hypermedia [Guéron et al. 2010]. Our aim is to develop a method for integrating models of cognitive science and empirical models extracted from traces of usages. The previous stages of this work resulted in a decision model for the task of Information Retrieval by Navigation, model which integrates some general strategies of information research and several cognitive styles. It also takes into account the position of links in the page and it can use a memory for some perceived objects.

These components are integrated in Virtual User (VU) but, we want to focus here on specific “knowledge”, i.e. associations between terms, extracted from navigational patterns. Our aim is to show the advantages they may offer. For the simulation, the task assigned to the VU is to find pages that match some keywords that are provided as input data (figure 10).

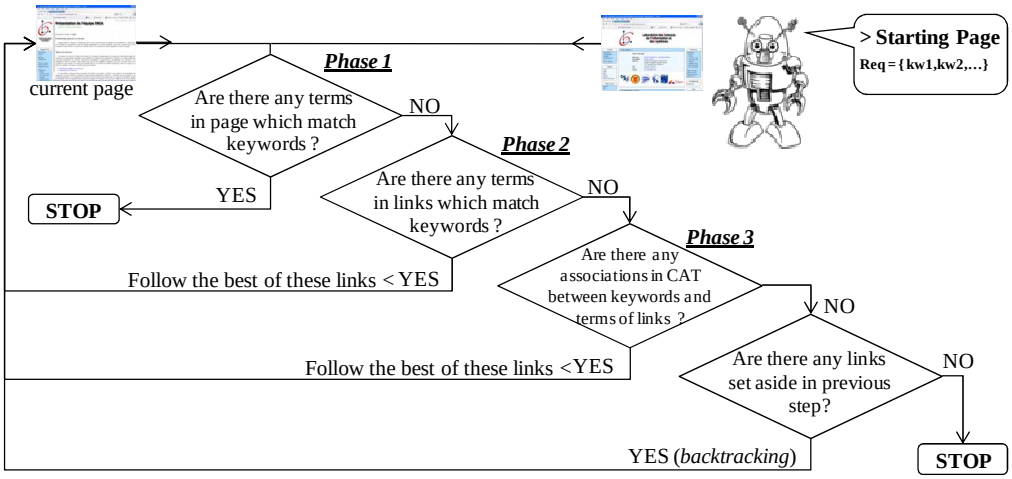


Figure 10. Behaviour of the robot "Virtual User" (*VU*).

As described in the figure above, the general algorithm governing the behaviour of the robot can be summarized in three main phases. The simulation starts with a given page and a given set of keywords. Then, the following three phases are repeated within a loop:

- ***Phase 1*** *:* Search for the needed information in the content of the current page: are there any terms visible on the current page matching the keywords ? If these terms exist, then the search will stop (success). Otherwise the robot continues with other phases ;
- ***Phase 2*** *:* Compare terms appearing on links of the current pages with keywords. If some terms of links match keywords select corresponding link ;
- ***Phase 3*** *:* Search in CAT for associations between the keywords and terms of visible links. Select these links if they exist.

The last phase is to return to the links not chosen in favour of better. The algorithm includes a lot of other particular cases which are not mentioned in these steps. Nevertheless, we do not pretend to reach a comprehensive model of the user in his/her task (conditions for ending, conditions for come back, strategies, styles of navigation …). Let's remind that the main issue here is to test the potential interest of the extracted associations. We can show the value-added of CAT associations using an example of navigation performed by the robot. We will describe its "journey" on two simple but significant examples. In order not to refer to specific elements of the site of our laboratory, we will use generic names for individuals and teams that appear in our examples. In particular, our tests focus on a researcher that we call *M. Smith* and a research team that we denote by *T_TEAM*.

Let's assume that the VU robot looks for information about *M. Smith* (its keywords) starting from the homepage of the web site (the starting page). First (phase 1, figure 11a), the robot will parse the contents of the homepage without success since the keyword {Smith} does not appear neither on this page nor on some visible links (phase 2). The figures 11a,b,c,d illustrate the different phases described in this paragraph.

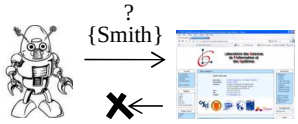


Figure 11a. Example of navigation

In such cases, a user would probably analyse the actual links on the page by performing cognitive operations to compare the terms of links with its own knowledge and, subsequently, to make a choice of actions. The robot will simulate such a process using not only some visible terms of links (phase 2) but also with CAT associations for taking advantage of other significant relations between terms (phase 3). That's the case here, for the association between *Smith* and *Directory (*figures 11b, 12a and 12b). If an association exists in both senses, we consider it more reliable than an unique sense association and it will be favoured.

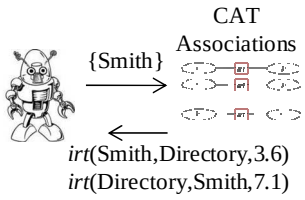


fig. 11b

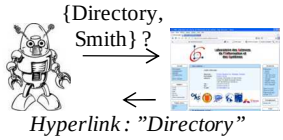


fig. 11c

fig. 11d

Figure 11 b,c,d. An example of navigation

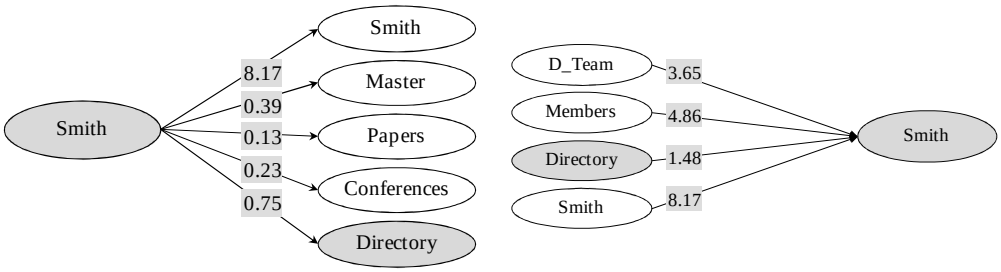
Then, the robot will search, among the terms displayed on the links of the page, those who are in relationships with keywords, according to the CAT associations (figure 11c). In our example, the VU will find a link with term *Directory*, visible for users.

Since a link showing the term *Directory* is visible on the current page, the robot will follow it towards the target page (figure 11d). In addition, the system adds the term *Directory* in the keywords list and re-started a new cycle (*Smith* is already in the set of keywords). By this way, we want to consider the evolution of the query during navigation based on information encountered by the user.

In this second cycle of navigation, the VU will find on the current new page (laboratory *Directory)* a hyperlink showing the term *Smith* (towards his personal page). This link is obviously chosen. As a result, the simulated navigation for the robot will be the following:

Laboratory Homepage → Directory → Smith Personal Page

We can see that having an association between *Smith* and *Directory* is **critical** since there is **not any information** in the homepage to guide users to information about *Smith*.



(12a) *(12b)*

Figure 12a and 12b. Weighted relations involving the keywords "*Smith*" in CAT associations

Let's consider another example with a request including two keywords *{Smith*; *D_Team}* and starting from the same page: *HomePage*. First, the *VU* which will not find any information in the current page (*HomePage*), will search for keywords on the hyperlinks. He will find a link entitled *D_TEAM* perfectly matching to the keyword *D_Team*, but nothing about *Smith*.

The *VU* saves this information in its working memory and continues to the next step. The robot's strategy is to seek first for links that satisfy all of the keywords. Otherwise the "direct links" (i.e. displaying terms of keywords) are first selected. Hence, the VU will look in *CAT* for any existing association with keyword *Smith*. It will found one between *Smith* and *D_Team* in both senses (Figure 12b presents one of them*)* and between *Smith* and *Directory*. The same process is carried out for the keyword *D_Team* but there isn't any association between *D_Team* and *Directory*. Consequently, due to the existing association between *D_Team* and *Smith*, the system will choose the hyperlink showing the *D_Team* term as the next to follow. The following is similar to the previous example. Arrived on the page *D_Team*, it will not find direct links with *Smith* but the CAT associations will highlight a link between *Member* and *Smith*. It will choose to follow this hyperlink towards a page where an explicit link to Smith's homepage exists. The final path of the VU will be the following:

HomePage → D_Team → Members → Smith Personal Page

Beyond these two examples we aim to show how the VU behaviour can be positively influenced by knowledge derived from associations. However, we wanted to add to these qualitative tests a more quantitative measure of the interest of contextual associations. We sought to show that for a significant number of sessions (not used for the learning phase of *CATMiner*), the paths followed by users can be given by some associations of *CAT*. We could treat each step of the pathway but with an aim of recommendation we focused on the first pages visited and the last reached pages. In other words the test can be verbalise as follows: is there a link between both according to one (ore more) association of *CAT* ?

To illustrate the principle and interest of such a test, let's take a simple example with the following session:

```
<internaute ident="5052" name="XX.XXX.XXX.XX" date="07/12/2008" start_time="19:29:00" end_time="19:30:15">
   <page name="Lab_Home_Page" id="6" time="19:29:00" />
   <page name="whos_who" id="25" time="19:29:12" />
   <page name="Personal_Page" id="1" time="19:29:21" />
   <page name="organisation" id="23" time="19:29:32" />
   <page name="Team" id="8" time="19:30:05" />
   <page name="Research_Topic" id="9" time="19:30:15" />
 </internaute>
```

Let's consider a user who would start his/her navigation from the homepage (*Lab_Home_Page*). He's searching for works dealing with *Modelling and Simulation*. He can't find explicit links on the homepage towards a page addressing this topic. However, this is an existing topic of a research team of the laboratory. Fortunately, an association between *D_Team* and *"Modelling and Simulation"* exists within the set of relationships extracted by *CATMiner*. It stands as an instance of the association *irt(Team Research_Topic)*: *irt(D_Team, "Modelling and Simulation").* Therefore, according to these relationships, the system will recommend to follow the link "*Team*", which is visible on the home page. On the page "*Team*" the user will found a link to the *Research Topic* of the team like for the last two pages of the session we present above. If we consider these two pages as being the user's goal, we can deem this recommendation as a positive one.

Through this experiment, we present some cases in which the links between terms (of *CAT*) permitted to reach the goal while the keywords and the information available on the pages would not enable the robot to take a logical decision. It is obvious that to demonstrate the usefulness of *CAT* associations, it would require to conduct large scale testing. The tests described below will allow us to produce statistically verified results.

4.2 Prediction of the Target Page

With this second experiment we aim to check whether, for a relatively large number of sessions, it is possible to provide such useful recommendation. Thanks to the log files, we have a relatively large number of sessions at disposal. For each session we consider the last page as being the target page of the user. We also consider the set of links, and associated terms, available on the current page from which the target page is looked for. And, of course, we use the set *CAT* of relationships extracted.

However, since the collection of traces is performed on server side, we do not have the keywords used, explicitly or not, by the user. Therefore we have to decide which keywords to use for representing the user's goal. We chose to adopt the terms appearing in the link of the last transition. If the goal is contained on the last page of the session we can assume that these terms have a strong link with the aim itself.

The figure below expresses for a given session *s*, the prediction test that we want to proceed on. Considering only the set $\{lt_4\}$, that is to say the words standing for the goal, and our associations *CAT*, we want to check whether it is possible to recommend the page p_5 as the next page to visit.

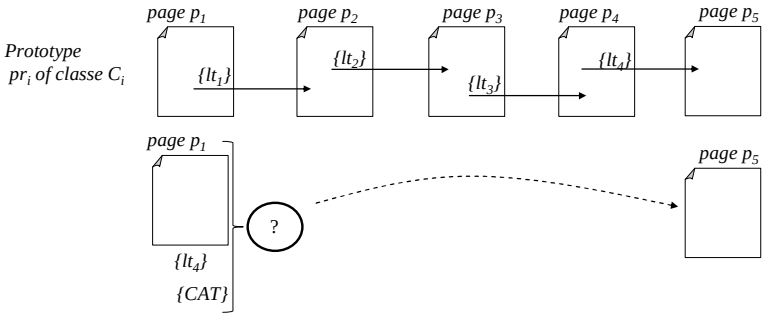


Figure 13. Principle of verification by early recommendation

The choice of keywords from terms of the last transition is an obvious bias of the experiment. But it seems realistic that this link has a strong relationship with the goal of the user. In addition, in *PatMiner* and *CatMiner* we only select sessions having at least 3 pages (the average is 5 pages per session). Therefore, it is likely that these terms do not appear on the start page of the session, which makes more challenging to be able to provide a relevant recommendation from these pages.

We have a set of 5000 sessions. About 4500 sessions, will be used to establish the *CAT* set of relations extracted by *CatMiner*. The prediction test is then performed on the 500 remaining sessions. The experiment will be repeated by changing the sessions that belong to the learning partition and the test partition according with the principle of the cross validation. The figure below shows the results of positive recommendations obtained for each of the 10 partitioning. The two rows correspond to predictions tests performed on the first page only (rank 1) or on the first and second page of the session (rank 2) when the system can't make a recommendation with the first page only.

With average rates of correct predictions above 65% at rank 1 and above 80% at rank 2, we can conclude that these results are positive.

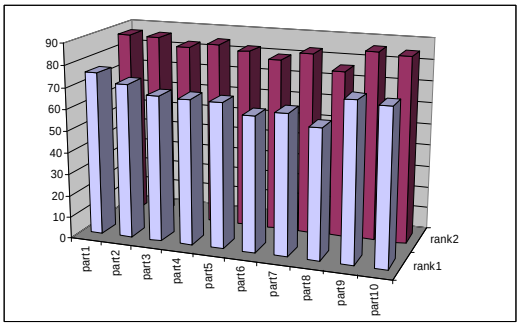


Figure 14. Percentage of positive prediction by using the CAT relations

This means that under certain conditions, the final searched page being may be anticipated without waiting for the 3rd, 4th, or even n-1th visited pages. But we also must put these results into perspective due to the choice of the keywords and also because of the type of web site we used. All of these features are potential biases of the experiment. However, these good values of current rates were obtained solely on the basis of contextual relations. Consequently, we can expect to conserve attractive values in less favourable conditions by using other networks of relationships between terms, as well as cognitive principles on strategies and trends in IR [Mroué & Caussanel 2009]. Therefore, in all cases we estimate to have a good capacity to guide the user in his/her task or to adapt information organisation to users' habits.

5. CONCLUSION

Improving the adaptation of the IS to its users requires a better understanding of the tasks performed. We must not only consider the tasks originally designed for the system but also the actual practices of users. The accurate observation of these practices can help to build some models of knowledge which can, in turn, help to guide users. In the work presented in this paper we sought to verify this chain of hypothesis.

From navigational patterns previously extracted, we considered the links followed by users and the terms that appear on these links. If we assume that there is a semantic link between these terms, we can, taking into account all classes of behaviour, develop a network of relationships between terms. At this stage we are not able to precise the nature of these relationships denoted in a generic way: *is-related-to.* Since these terms are related to the specific content of IS (as the name of a project, a product, a company, ...) the extracted relationships appear to be highly contextual. Although these are simple associations, and covering a relatively small number of terms, we have shown that these relationships provide sufficient information to predict user behaviour under certain conditions. We explain such effectiveness by the fact that the terms and the relations come from actual practices occurring in the web site.

Traces of navigation provide a limited view (without argumentation …) of user activity but with huge advantage to describe actual behaviour. The challenge is therefore to give sense to the raw data to enrich positive cognitive models. Often considered as inoperable, these traces seem to us underutilized in models of user. Our goal was primarily to show the relevance of knowledge acquired from these traces for decision models and more generally the relevance of

such an inductive approach. That's what we did and presented here, with results that confirm the interest to extract and use such contextual associations in order to support users in their everyday tasks.

These positive results allow us to consider in a very short term new developments and improvements: using blocks of text more than just words of links, by addressing images links, by asking human subjects to verbalize their goal before starting their research and to justify their choices during their navigation, etc.Others immediate perspectives deals with the extensions of the method to another type of site, to another type of environment and to integrate some cognitive models developed to provide an adaptation even more focused.

REFERENCES

[Anderson et al. 2004] Anderson, J. R., Bothell, D., Byrne, M.,Douglass, D., Lebiere, C., & Qin, Y., 2004, 'An integrated theory of mind', *Psychological Review*, 111, 1036–1060.

[Baeza-Yates & Tiberi 2007] Baeza-Yates R., Tiberi a., 2007, 'Extracting semantic relations from query logs', in the proceedings of the *13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, California, USA, 76 – 85.

[Benslimane et al. 2006] Benslimane D., Arara A., Falquet G., Maamar Z., Thiran P., Gargouri F., 2006, 'Contextual Ontologies', in Proceedings of the *Fourth Biennial International Conference on Advances in Information Systems*, 18-20 October, 2006 Izmir, Turkey.

[Bouquet et al. 2003] Bouquet P., Giunchiglia F., Van Harmelen F., Serafini L., Stuckenschmidt H., 2003, 'C-OWL: Contextualizing ontologies', *Journal Of Web Semantics*, 164--179, Springer Verlag.

[Campbell et al. 2007] Campbell D. R., Culley S. J., McMahon C. A. , Sellini F., 2007, 'An approach for the capture of context-dependent document relationships extracted from Bayesian analysis of users' interactions with information', *Information Retrieval*, Volume 10, Number 2, Pages 115-141.

[Chaker et al. 2010] Chaker H., Chevalier M., Soulé-Dupuy C., Tricot A., 2010, 'Improving information retrieval by modelling business context', in *International Workshop on User Profiles in Multi-application Environments*, at CENTRIC 2010, p. 121-126, Nice,France.

[Conde et al. 2010] Conde J.M., Vallet D, Castells P., 2010, 'Inferring user intent in web search by exploiting social annotations', in Proceeding of the *33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, 827—828, Geneva, Switzerland.

[Cool & Spink 2002] Cool C., Spink A., 2002, 'Issues of context in information retrieval (IR): an introduction to the special issue', *Information Processing and Management*, 38, 5, 605-611.

[Dinet & Tricot 2008] Dinet, J., Tricot, A., 2008, 'Recherche d'information dans les documents électroniques', in A. Chevalier & A. Tricot, (Eds.), *Ergonomie des documents électroniques*, pp. 35-69, Paris : PUF.

[Eirinaki et al. 2004] Eirinaki, M., Lampos, C., Paulakis, S., and Vazirgiannis, M., 2004, 'Web personalization integrating content semantics and navigational patterns', in Proceedings of the *6th Annual ACM international Workshop on Web information and Data Management* (Washington DC, USA, November 12 - 13, 2004). WIDM '04. ACM, New York, NY, 72-79.

[Fu & Pirolli 2007] Fu, W., Pirolli, P. L., 2007, 'SNIF-ACT: a cognitive model of user navigation on the World Wide Web', *Human Computer Interaction*, vol 22 (4): 355-412.

[Guéron et al. 2010] Guéron D., Maille Nicolas, Caussanel J., Chaudron L., 2010, 'Aggregation of Human Activities, Building Flight Profiles', poster of *International Conference DOM-VI, Decade of the Mind*, "Looking forward to the next ten years". Singapore 18-20 October.

[Herder 2006] Herder E., 2006, 'Forward, Back and Home Again, Analyszing User Behavior On The Web', *Doctoral dissertation*, University of Twente.

[Hernandez et al. 2007] Hernandez N., Mothe J., Chrisment C., Egret D., 2007, 'Modeling context through domain ontologies', *Information Retrieval, 2007, Volume 10, Number 2, Pages 143-172*, Volume 10 Issue 2.

[ISO/IEC 1998] International Organization for Standardization, ISO/IEC 13250, Information Technology-SGML Applications-Topic Maps, Geneva: ISO, 1998.

[Juvina 2006] Juvina, I., "Development of a Cognitive Model for Navigating on the Web", 2006, *Doctoral dissertation at Utrecht University*.

[Kitajima et al. 2000] Kitajima, M., Blackmon, M.H., Polson, P.G., 2000. 'A Comprehension-based Model of Web Navigation and Its Application to Web Usability Analysis', In S. McDonald, Y. Waern & G. Cockton (eds.), People and Computers XIV - Usability or Else! *Proceedings of HCI 2000*, Springer, pp.357-373.

[Landauer & Dumais 1997] Landauer, T. K., Dumais, S. T., 1997, 'A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge', *Psychological Review*, 104, 211-240.

[Liu & Chu 2007] Liu, Z., Chu, W.W., 2007, 'Knowledge-based query expansion to support scenario-specific retrieval of medical free text'. *Information Retrieval*, 10, 2, 173-202.

[Mroué & Caussanel 2010] Mroué A., Caussanel J., 2010, 'Traces of navigation as knowledge for simulating Information Retrieval by navigation task in hypertext document', *in the fourth international IEEE conference RCIS, Research Challenge in information Science*, 19-21 May, Nice, France.

[Mroué & Caussanel 2009] Mroué A., Caussanel J., 2009, Anticipate site browsing to anticipate the need, in: Springer (Ed.), *Web Mining Applications in E-commerce and E-services Series: Studies in Computational Intelligence*, I-Hsien T., Hui-Ju W. eds, vol. 172. ISBN: 978-3-540-88080-6

[Mylonas et al. 2008] Mylonas P., Vallet D., Castells P., Fernández M., Avrithis Y.S., 2008, 'Personalized information retrieval based on context and ontological knowledge', *The Knowledge Engineering Review*, 23, 1, 73-100.

[Pirolli & Card 1995] Pirolli P., Card S., 1995, 'Information Foraging in Information Access Environments', in Proceedings of the *Human Factors in Computing Systems, CHI '95*. Association for Computing Machinery.

[Polaillon et al. 2007] Polaillon G., Aufaure M-A., Le Grand B., Soto M., 2007, 'FCA for contextual semantic navigation and information retrieval in heterogeneous information systems', in proceedings of the *18th International Conference on Database and Expert Systems Applications*, p.534-539.

[Santos & Nguyen 2009] Santos E. Jr., Nguyen H., 2009, 'Modeling Users for Adaptive Information Retrieval by Capturing User Intent', *Collaborative and Social Information Retrieval and Access: Techniques for Improved User Modeling*, Edited by M. Chevalier, C. Julien, and C. Soule-Depuy, 88-118, IGI Global.

[Sondhi & Chandrasekar 2010] Sondhi P., Chandrasekar R., 2010, 'Domain Specific Entity and Relationship Extraction from Query Logs', in Proceeding of *ASIS&T 2010*, Oct 2010, Pittsburgh, USA.

[Tamine-Lechani et al. 2010] Lynda Tamine-Lechani, Mohand Boughanem and Mariam Daoud, Evaluation of contextual information retrieval effectiveness: overview of issues and research, *Knowledge and Information Systems, 2010, Volume 24, Number 1, Pages 1-34.*

[Tricot & Nanard 1998] Tricot, A., Nanard, J., 1998, 'Un point sur la modélisation des tâches de recherche d'informations dans le domaine des hypermédias', *Hypertextes et Hypermédias*, n° hors série, 35-56, 1998.

[Van Oostendorp & Juvina 2007] van Oostendorp H., Juvina I., 2007, 'Using a cognitive model to generate web navigation support', *International Journal of Human-Computer Studies*, 65 ,10, 887-897.

[XTM 2006] XML Topic Maps (XTM) 2.0, Topic Maps — XML Syntax, http://www.isotopicmaps.org/sam/sam-xtm/, 2006.

[Xu et al. 2008] Xu S., Bao S., Fei B., Su Z., Yu Y., 2008, 'Exploring folksonomy for personalized search', in Proceedings of the *31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, 155—162, Singapore.