Query Reformulation for Information Retrieval on the Web Using the Point of View Methodology: Preliminary Results

Leila Naït-Baha, Agata Jackiewicz, Brahim Djioua, Philippe Laublet

Langage, Logique, Informatique et Cognition team (LALIC)
Centre d'Analyse et de Mathématiques Sociales (CAMS)
UMR 8557 du CNRS, EHESS, Université de Paris-Sorbonne
96, Boulevard Raspail, 75006 PARIS FRANCE
Tél.: (33) 01 44 39 89 63 Fax: (33) 01 44 39 89 68
naitbaha@caramail.com, ajackiewicz@paris4.sorbonne.fr,
bdjioua@paris4.sorbonne.fr, Philippe.Laublet@lip6.fr

Leila Naït-Baha is a post-graduate student at the Human Science Institute, Paris-IV Sorbonne University, France. Her professional interests lie in the characterisation of users' informational needs regarding textual information retrieval on the Web.

Philippe Laublet is a professor at the University of Paris-Sorbonne. Prior to that, he was senior researcher at the French National Research Center for Aeronautics and Space. His major areas of research are knowledge engineering, knowledge acquisition, object-oriented methods and the semantic web.

Leila Naït-Baha, Agata Jackiewicz, Brahim Djioua, Philippe Laublet. (2001). **Query Reformulation for Information Retrieval on the Web Using the Point of View Methodology: Preliminary Results.** *Knowledge Organization, 28(3).* 129-136. 10 refs.

ABSTRACT: The work we are presenting is devoted to the information collected on the WWW. By the term collected we mean the whole process of retrieving, extracting and presenting results to the user. This research is part of the RAP (Research, Analyze, Propose) project in which we propose to combine two methods: (i) query reformulation using linguistic markers according to a given point of view; and (ii) text semantic analysis by means of contextual exploration results (Desclés, 1991). The general project architecture describing the interactions between the users, the RAP system and the WWW search engines is presented in Naït-Baha et al. (1998). We will focus this paper on showing how we use linguistic markers to reformulate the queries according to a given point of view.

KEY WORDS: information retrieval on the Web, RAP project, query reformulation, point of view, linguistic markers, causality.

1. Introduction

This work is dedicated to information retrieval on the Web and is part of the RAP[1] project "Research, Analysis and Proposal". This project presents the combination of two methodologies: i) query reformulation using linguistic markers that correspond to a given point of view and ii) text semantic analysis by means of contextual exploration results (Desclés, 1991). In this paper we will mainly focus on reformulation issues, the entire project having already been presented in Naït-Baha *et al*. (1998).

2. Issues

The usual procedure for an information retrieval session consists of elaborating a query from nominal information referring to topics implied by the search,

as in the following example: (i) "migraine", (ii) "migraine" and "long term treatment" or "drugs for crisis". The better the user knows the domain of investigation and its associated specific terminology, as well as the unquestionable identifiers associated with the object of the search, the easier it is for the user to follow this procedure. The user can then easily refine the search if the search engine brings back a noisy amount of replies. But when information retrieval is associated with an emerging domain, or a domain which is not familiar to the user, or when information retrieval is not guided by a precise and clearly identifiable topic, query formulation then becomes more hazardous for the user. At the same time, the use of Boolean queries, even improved with the use of various operators such as truncation or proximity operators requires the use of a specific formulation – which is not always done. It has been observed by G. Grefenstette (1997) that most queries elaborated by web users who are not documentation specialists are either made of very few words[2], or their reformulation generally consists of (i) repeating the initial query, (ii) adding or deleting a few words, (iii) spelling the query differently, or (iv) using derived forms or abbreviations. Users also associate a set of synonyms or related concepts to the words of the initial query (Bruza & Dennis, 1997). More generally, it has been noticed that the final query is only a partial indicator of the user needs in terms of information (Le Coadic, 1998).

In order to assist the user to complete exploratory searches and specify his or her needs, we propose that the search area be divided into several sub-areas materialized by specific points of view. Each point of view is intended to capture a specific informational trace. This method will highlight some of the kinds of information contained in a web page. Then extraction of textual information is carried out on the selected documents (after a reformulation of the initial query according to the stated points of view), using the semantic analysis performed by the FILTEXT software platform (Minel et al., 1999). The first step of the processing involves associating a set of linguistic markers that correspond to each of the selected viewpoints to the terms of the user query. Then the queries are automatically reformulated in a preprocessing operation, according to the selected viewpoints and the dynamic integration into the reformulated query, which is a set of markers that target the information that should be selected in the documents. We make use of cross-domain reformulations with regard to the explored domains, which means that the elaboration of these markers does not rely on the knowledge associated to these particular domains.

The software prototype of the RAP system has been initiated by Brahim Djioua, and then taken up by Leïla Naït-Baha (Djioua *et al.*, 1998).

3. Linguistic points of view in RAP: the example of causality

The unique purpose of our project was to take the necessary steps to access the semantic information contained in the texts, in order to specify and extract the most relevant sequences. For this reason we chose to make use of a purely linguistic knowledge and, more specifically, a semantic knowledge. We will focus on the textual expression of some knowledge organizational conceptual relations, such as causal and definitional relations, and also some static relations[3]. At the present time, only the causal point of view is available in RAP. Causal knowledge is defined in Agata Jackiewicz's work (Jackiewicz, 1998). The points of view relating to definitions and static relations are still under study.

Causal relations allow the exploration of the information contained in a specific domain and its subdomains, and also create a link between several domains of knowledge, providing cross explanations, in a varied and pluridisciplinary approach of reality. They often reveal unexpected or even surprising associations between objects or situations. These unexpected associations are why the filtering of causal utterance applies to an exploratory type of information retrieval oriented towards the gathering of new types of information. To discover unexpected links between apparently well-defined and usually not interdependent domains is one of the main motivations for text mining. We will analyze the textual expression of causality by distinguishing four different approaches to this notion. Each of these approaches can be identified through the texts by a set of specific markers (as in table no. 2). The **qualitative** approach corresponds to the most widely accepted approach of causality. Being both intuitive and rich in linguistic expression, it has a strong explanatory power and clearly distinguishes the cause of an event from its consequences (as in example no. 1 below). The **functional** approach provides the reliability and coherence of statistic measures that can either reveal or confirm the existence of a causal relation. It applies to big corpora on which it aims at establishing general rules (as in example no. 2). The **analytical** approach of causality brings to light

some factors that can only be efficient from a causal point of view with regards to a produced effect or consequence. Its partial aspect either reveals partial indetermination with regards to the implied factors or a focus of the interest on selected specific factors (as in example no. 3). The **synthetic** approach reveals complex or extended links that can be expressed generally, but that must be made explicit to prove their existence (as in example no. 4 below).

(Example 1) *Le mariage **est l'aboutissement de** plusieurs actes élémentaires du rituel de la cour, voire de plusieurs cours successives, et les délais de mariage dans une génération sont distribués selon une loi gamma.(H. Le Bras, Pour la science, avril 96).*
(Example 2) *Le nombre de conjoints potentiels ne **dépend** pas seulement **de** paramètres objectifs de lieu, de fortune ou de capacité, mais aussi du temps qu'on passe à faire la cours. (H. Le Bras, Pour la science, avril 96).*
(Example 3) *Qui peut être certain, en effet, que l'absorption à haute dose des niaiseries qui constituent 90 % des programmes n'**aura** pas, à moyen ou long terme, un **effet délétère irréversible sur** la conscience même des téléspectateurs ? (LMD, juin 96, p.32)*
(Example 4) *Le **lien entre** la filiation matrilinéaire **et** le matriarcat ne survécut pas très longtemps aux observations. En 1915, déjà, W.H.R. Rivers "déconstruisait" soigneusement la notion de "droit maternel", montrant qu'**il n'y avait** pas de **lien entre** le régime de filiation **et** la position occupée par la mère dans le foyer ni, a fortiori, par les femmes dans la société tout entière.*

The markers corresponding to these approaches of causality, can be used to reformulate queries that match the expectations of an extended set of users (neophyte, information miner, researcher). As a matter of example, a thorough study of a multiple causal determination phenomenon, such as the green house effect, can rely more specifically on the analytical approach of causality. It appears in textual sequences through the significant presence of markers such as *contribution, contribuer à, la part de dans, compter pour dans* (as in example no. 5).

(Example 5) *Quels sont donc ces gaz dont **les contributions** respectives **à** l'effet de serre vont évoluer rapidement lors des prochaines années? – Le dioxyde de carbone (CO2) **compte pour** 49 % actuellement **dans** ce phénomène, mais ne représentera plus que 40 % d'ici une dizaine d'années. – Les chlorofluorocarbones (CFC), également responsables de la destruction de l'ozone stratosphérique en Antarctique, **contribuent au** réchauffement climatique à hauteur de 20 %, et bientôt 25 %. – Le méthane (CH4), dont la concentration augmente de 1 % par an et **dont la part dans** le réchauffement **se situe** pour le moment **vers** 12 %. L'oxyde nitreux (NO2), **responsable de** 10 % environ de l'effet de serre. (LMD, fév. 90, p.27)*

4. General architecture of the RAP system

The RAP system can be set to two different operating modes: the "administration" mode, and the "user" mode. The "administration" mode allows for the management of the linguistic markers associated with the user's search points of view, whereas the "user" mode allows for the typing of the initial query, and the choice of a search environment (criteria and points of view), together with the processing of the search and viewing of the results.

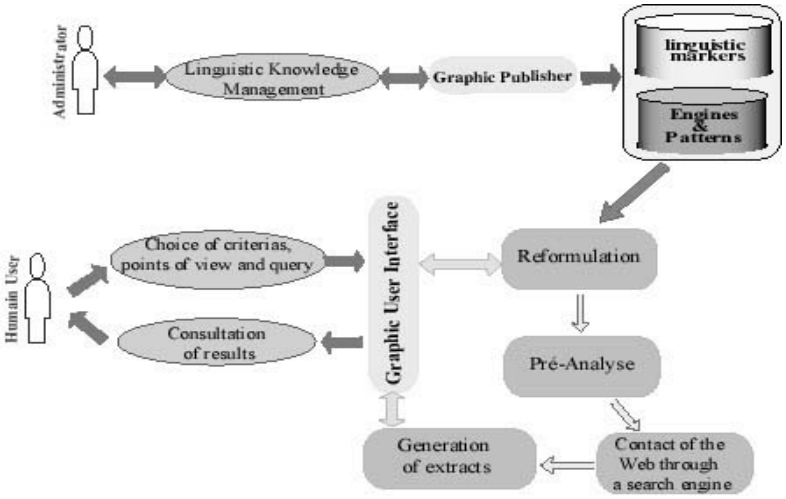

Figure 1: RAP system architecture

The RAP system administration interface can be used to manage the set of linguistic markers associated with a specific point of view. This module is the RAP system feedback interface. Session after session this module allows for the management and readjustment of the most relevant markers for a given point of view. The prototype version of RAP comes along with a Java applet that allows users to connect to the RAP architecture using a web browser (such as Netscape), in the same way they would use a classical search engine. Launching the query will generate as many reformulations as there are selected points of view. Keyword query reformulation makes use of patterns associated to the specific syntax of the search engine used (here AltaVista™).

The pre-analysis stage input data consists of web page addresses sent back by the search engine that are considered relevant to the user's specific criteria. At this stage the RAP system automatically eliminates the addresses that do not match the user's chosen points of view (some criteria such as last modification date are immediately taken into account when the initial query is reformulated). The pre-analysis operation output is a list of the URL addresses of the pages that will be submitted to semantic analysis. The semantic analysis module "ANASEM"'s input parameters before completion of the semantic analysis are the HTML pages associated to these addresses together with the user's points of view. The purpose of this stage is to generate a set of text extracts relevant to the user's points of view. The user can then view the HTML format extracts sent back by the ANASEM module. When users judge the extracts relevant they can then browse the original web site of this page using a web browser. It is important to note however, that at the present time, paragraph extraction is based on resource saving heuristics, the relevance of which we are still testing. These heuristics consist of taking into account the density of the chosen point of view markers, through the different paragraphs of a single web page. This allows the selection of the very "best" paragraphs of a set of "n" pages ranked first by the search engine after they have been reformulated, and then the presentation of the selected paragraphs to the user in the relevant format.

5. Reformulation in RAP

In many existing systems (Bruza & Dennis, 1997; Van Der Pol, 1996; Nie *et al*., 1997), reformulations are made at either a pre-treatment or post-treatment stage. They can rely on either search domain hierarchy or the use of semantic relations based on specific terminology, or even on the creation of tools that provide a dynamic extension of the query. In the RAP system, reformulation is the result of a pre-treatment dynamic operation. It requires user interaction to make a certain number of choices such as: (i) the initial query; (ii) the selection criteria (document type, last modification date, *etc*.) that will allow a pre-filtering of the information among the number of documents that might possibly answer the initial query; (iii) points of view that will focus the search on specific domains according to a well determined direction and, finally, (iv) the reformulation modes. Then, the RAP system sends the reformulated query to a search engine for treatment (AltaVista™ for the present time). After project finalization the system should offer users the option to choose their favorite search engines.

We have elaborated three distinct modes of reformulation: the *extended* mode, the *targeted* mode, and the *targeted restricted* mode. Each mode puts specific constraints on search clause locations within the HTML pages (body, title, etc.), and a proximity constraint (distance between words) that applies to the distance between the initial query and the set of point of view markers selected by the user. In the *extended* mode, proximity between the initial user query and the whole set of user specific point of view markers is searched for in any part of the HTML documents (body, title, URL, anchor, *etc*.). In the *targeted* mode only the HTML documents whose titles match the initial user query are taken into consideration. Afterwards, proximity between the initial user query and whole set of point of view markers is then searched for in any part of the already pre-selected documents (body, title, url, anchor, *etc*.). In the *restricted targeted* mode only those web pages that match the proximity constraint between the initial query and the markers, with the position of the title within the HTML documents are selected.

As already stated, each specific point of view is associated with a set of linguistic markers that express the point of view in written documents in a one-to-one relation and repeated manner. More precisely, the selection of markers for reformulation purposes has been made in order to satisfy the three following constraints: (i) identify the most frequent markers, the size of the reformulated query being limited (to 260 characters on AltaVista™), meaning that only about twenty markers can be associated to each point of

Reformulation Modes	Reformulation patterns	Examples
extended	(X ***NEAR*** Y)	(dioxine ***NEAR*** (provoq* ***OR*** caus* ***OR*** *entraîn** ***OR*** ...))
targeted	(***title*** :X ***AND*** (X ***NEAR*** Y))	(***title*** :dioxine ***AND*** (dioxine ***NEAR*** (provoq* ***OR*** caus* ***OR*** *entraîn**...)))
restricted targeted	(***title*** :X ***NEAR*** Y)	(***title*** : dioxine ***NEAR*** (provoq* ***OR*** caus* ***OR*** *entraîn**...))

Table no. 1: Reformulation modes and their respective patterns (X stands for the initial user query, Y for the logical disjunction between makers related to a chosen point of view; here, causality).

Approach	Association markers for reformulation phase
Qualitative (P1)	*provoqu*, caus*, déclench*, entraîn*, détermin*, émerg*, engendr*, indui*, occasionn*, effet, conséquence, origine, source, résult*, favoris*, dû, issu, modifi*, augment*, diminu**
Analytic (P2)	*contribu*, impact, influ*, incidence, particip*, rôle, intervention, emprise, poids dans, retentissement, action sur, conséquence sur, pouvoir sur, effet sur, rôle dans, en cause dans, impliqué dans, part à, implication dans, responsabilité dans,*
Functional (P3)	*corrélation, corrélatif, corrélativement, en fonction de, rapport mutuel, interdépendance, lien réciproque, covariance, covariation, relation réciproque, relation linéaire, évolution parallèle, dépendance mutuelle, fonction linéaire, sous la dépendance de, à mesure que*
Synthetic (P4)	*lien, relation, rapport, rapprochement, est lié, sont lié*, être lié*, vont de pair, vont de concert, vont de conserve, relier à, lier à*

Table no. 2: Linguistic French markers associated to distinct approaches of causality.

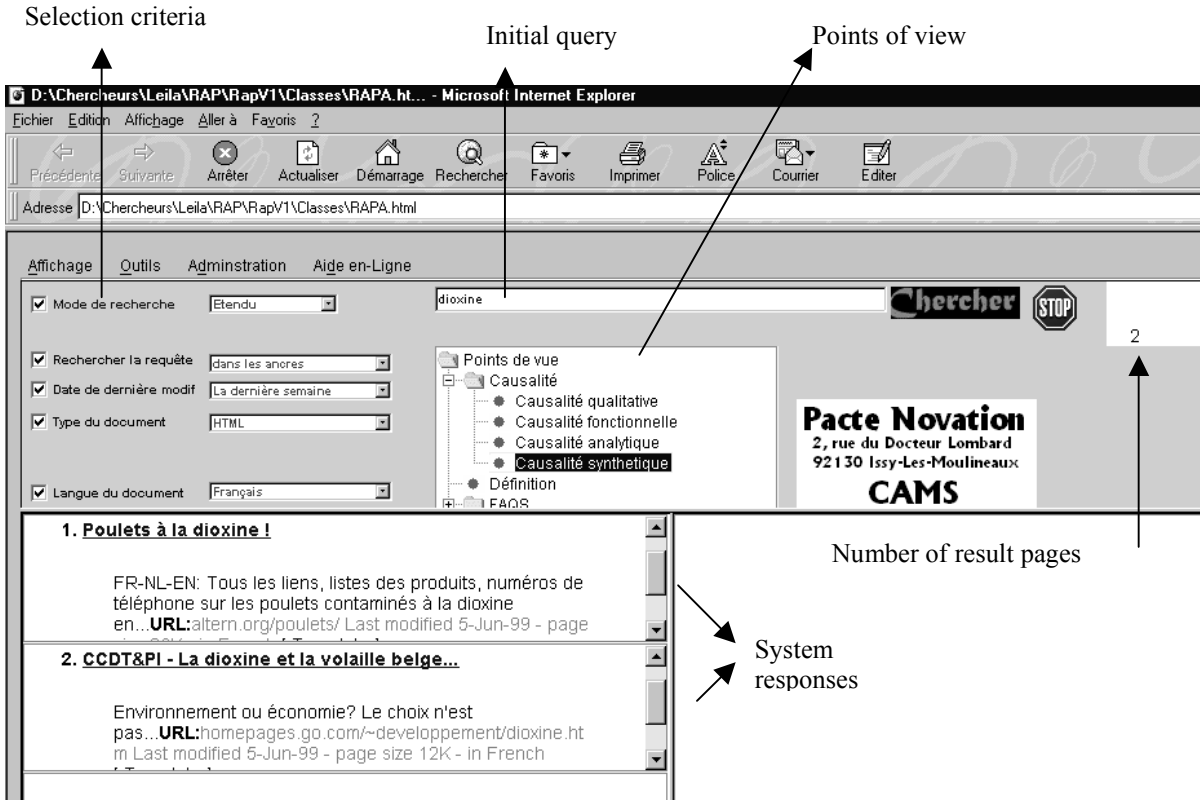

Figure 2: Processing of the query dioxine *choosing the following point of view:* synthetic causality

view; (ii) choose the most reliable markers (i.e. the least polysemic); (iii) prefer simple markers to compound markers (*cause of* will be preferred to *constitutes the cause of*); the problem of verb inflected forms is resolved by using the radical followed by the truncation symbol "*" (*provok**). It should be noted though that the selection of relevant extracts implies the use of an extended set of sub-markers. Our data-

base presently consists of a set of 1000 causality markers.

6. Information retrieval with RAP: examples

Figure no. 2 shows the user interface proposed for an information retrieval session with the RAP system.

The causal point of view is structured into four distinct perspectives that correspond to the four different approaches i.e., qualitative, functional, analytical and synthetic. A summarized presentation of each approach (such as definition, sample phrase and list of most significant markers) for illustration purposes is made in the system online help.

We have carried out a total of fifty test sessions: in French, with and without reformulation, with and without result ranking, using a combination of the three reformulation modes and the four causal approaches. The date of the last document modification was set between 1/1/1996 and 12/1/1996. All these elements were taken into account for the dynamic reformulation of the initial query. Results were ranked by AltaVista™, taking into account the initial query along with the markers associated to the four different approaches of causality. The tables below are a summary of the results obtained with the initial query "el niño".

Initial query = "el nino" Language = *french* Date = 1/1/96- > 31/12/99 With document ranking	**without Reformulation** search engine response: 1180 Web pages			
with Reformulation	*P1 (Qualitative)*	*P2 (Analytic)*	*P3 (Functional)*	*P4 (Synthetic)*
extended mode	137	153	65	62
targeted mode	12	14	4	4
restricted targeted mode	0	0	0	0

Table no. 3: Number of pages sent back using the query: "el niño" with and without reformulation

Tables 4 and 5 show the five first search engine replies in relevance ranking

Without Reformulation
1. **El Nino** [**El niño** L'enfant Jésus. Le phénomène **El Niño** est une catastrophe climatique assez particulière.
2. **El Niño** risque de réduire les approvisionnement vivriers en ... [nbsp; **El Niño** risque de réduire les approvisionnements vivriers en Asie. Les conditions de sécheresse extrême menaçant la sécurité alimentaire dans...]
3. L'épisode **El Niño** que connait le pacifique tropical pourrait bien devenir l'ev... [wmo, omm, geneva, switzerland]
4. **El Niño** [Description du phénomène **El Niño**]
5. **El NINO** [Yahoo! France. Yahoo! France Résultat de la recherche : Yahoo! Annuaires – Actualité – Nouveautés – ...

Table no. 1: Five first results after processing of the query "el niño" without reformulation

With Reformulation(targeted model, qualitative causality)
1. Actualités : FAO, **El Niño** ***provoque*** des dégâts considérables ...
2. **El Niño** [Description du phénomène **El Niño**]
3. La Recherche – Heurs et malheurs de la prévision d'**El Niño** [La Recherche est la première revue internationale d'information scientifique et technique de langue française]
4. **El Niño** [**El Niño**: Un phénomène climatique exceptionnel. ***Origine***. ***Influence***. Prévisions. ***Impact sur*** la flore marine.
5. **El Niño** risque de réduire les approvisionnement vivriers en ... [nbsp; **El Niño** risque de réduire les approvisionnements vivriers en Asie. Les conditions de sécheresse extrême menaçant la sécurité alimentaire dans...]

Table no. 2: Five first results after processing of the query "el niño" with reformulation

We would be very eager to present the following extract, with a ranking of the most relevant paragraphs (based on the criterion of marker density according to the chosen point of view), to the user. The extraction rule used consists of considering marker density for the chosen point of view on the distinct paragraphs of a single web page. It allows to select the very "best" paragraphs of a set of "n" pages ranked first by the search engine after they have been reformulated. The selected paragraphs are then presented to the user in the relevant format.

> EL NIÑO A ***PROVOQUE*** DES DEGATS CONSIDERABLES A L'AGRICULTURE, AUX FORETS ET AUX PECHERIES
> "Rome, 31 juillet 1998.- Le phénomène météorologique **El Niño**[4] ***a provoqué*** des dégâts considérables à l'agriculture, aux forêts et auxpêcheries, plus particulièrement en Asie et en Amérique centrale et latine, souligne un rapport publié aujourd'hui par l'Organisationdes Nations unies pour l'alimentation et l'agriculture (FAO). Des inondations ont été signalées dans 41 pays. Vingt-deux pays ont été

affectés par la sécheresse et deux autres par des incendies de forêts particulièrement graves. Outre ses ***effets*** à long terme ***sur*** la production agricole et la situation alimentaire de ces pays, **El Niño** a également ***entraîné*** des conditions propices à des épidémies et à des épizooties graves. Ainsi, la fièvre dite de la vallée du Rift a dévasté plusieurs régions au Kenya, en Somalie et en Ethiopie. En ce qui concerne les pêches, **El-Niño** *a* eu un ***impact*** considérable au large des côtes ouest de l'Amérique du Sud, une des régions les plus riches en poisson de la planète (12 à 20 pour cent des débarquements mondiaux). La production d'anchois pourrait s'en trouver ***affectée*** du fait de la baisse des stocks au large du Pérou et du Chili. Les incendies de forêt ***provoqués*** par **El Niño** ont, par ailleurs, des ***incidences*** non seulement sur la sécurité alimentaire mais aussi sur le réchauffement climatique ***dû*** à l'***effet*** de serre. Le Système mondial d'information et d'alerte rapide (Smiar) de la FAO, dirigé par Abdur Rashid, suit en permanence les ***effets*** d'**El Niño** ***sur*** les cultures et la situation des approvisionnements alimentaires dans diverses parties du globe et publie des mises à jour périodiques." http://www.waternunc.com/fr/fao3.htm

7. Conclusion

The first version of the RAP prototype is currently being evaluated at URFIST in Paris by Ghislaine Chartron's team. But many aspects of the project are still under development. The integration to the RAP system of an extended set of points of view (definition, quote, static relations, *etc.*) is still under study. We are also considering the gradual building of an associated set of points of view (causality, definition, quote, static relations, *etc.*) in order to answer increasingly complex needs that can be shared by a wide community of users. In addition, we are working to produce a finer grained analysis for the selection of relevant extracts by comparing the results obtained with the two methods detailed above: the first one relying on a thorough but resource consuming semantic analysis through contextual exploration, and the second one on the use of marker density in paragraphs and their distance with respect to the terms of the initial query.

8. Notes

1 The CAMS' partners for this project are "Pacte Novation" and URFIST in Paris. The *RAP* project is financed by the French Ministry of Research, Public and Higher Education and and the ministry of Industry, together with the Ministry of Post and Telecommunications. The coordonator for the RAP project is Philippe Laublet, working under scientific responsibility of Jean-Pierre Desclés.

2 Querying a search engine on a meteorological phenomenon such as El Niño with the simple query "el nino" has the effect of taking the risk to browse a considerable amount of noisy information: 73314 documents (all languages) and 745 documents (French). Searches have been carried out on May 15th 1998 on Alta Vista™.

3 Spatial relations: part-whole relation, inclusion relation, relevant relation...

4 Le gras marque la requête initiale, l'italique représente les marqueurs du point de vue choisi.

9. Bibliography

Bruza, P.D. & Dennis, S., (1997). *Query Reformulation on Internet: Empirical Data and the Hyperindex Search Engine*.

Desclés, J.-P., Jouis, C., Oh, H.-G. & Reppert, D., (1991). "Exploration contextuelle et sémantique: Un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte", in *Knowledge modeling and expertise transfert*, D. Hérin-Aime, R. Dieng, J.P. Regourd, J.P.Angoujard (éds), Amsterdam, Washington DC, Tokyo : IOS Press.

Djioua, B., Jackiewicz, A., Laublet, P. & Naït-Baha, L. (1998). "Recherche et filtrage d'information dur les réseaux", *Rapport de Recherche du projet RAP pour le Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche*.

Grefenstette, G. (1997). "Short Query Linguistic Expansion Techniques: Palliating One-Word Queries by Providing Intermediate Structure to Text", in M.T. Pazienza (Ed.) *Information Extraction, Springer*.

Jackiewicz, A. (1998). *L'expression de la causalité dans les textes. Contribution au filtrage sémantique par une méthode informatique d'exploration contextuelle*. Thèse de doctorat. Université de Paris-Sorbonne, 1998.

Le Coadic, Y.F. (1998). "Le besoin d'information, formulation, négociation, diagnostic", Paris: ADBS Editions.

Minel, J.-L., Desclès, J.-P., Cartier, E., Crispino, G., Ben Hazez, S. & Jackiewicz, A. (1999). "Résumé automatique par filtrage sémantique dans les tex-

tes. Présentation de la plate-forme FilText". *Revue Technique et Science Informatique*, no. 3, 2001

Naït-Baha, L., Jackiewicz, A. & Laublet, P (1998). "Reformulation de requêtes et extraction de phrases pertinentes pour la collecte d'informations sur le Web". *In Proceedings of RIFRA'98, Sfax*, Tunisie.

Nie, J., Chevallet, J.P., Chiaramella, Y. (1997). "Vers la recherche d'informations à base de termes". *In Proceedings of the 1st JST FRANCIL de l'AUPELF*, Avignon, France.

Van der Pol, R.W. (1996). *A device for query composition*, Report CS96-02, Université de Maastricht.