Analyzing Temporal Query for Improving Web Search

Rim Faiz
LARODEC, IHEC, University of Carthage, Tunisia
E-mail: Rim.Faiz@ihec.rnu.tn

***Abstract*— Research of pertinent information on the web is a recent concern of Information Society. Processing based on statistics is no longer enough to handle (i.e. *to search, translate, summarize...*) relevant information from texts. The problem is how to extract knowledge taking into account document contents as well as the context of the query. Requests looking for events taking place during a certain time period (i.e. *between 1990 and 2001*) cannot provide yet the expected results. We propose a method to transform the query in order to "understand" its context and its temporal framework. Our method is validated by the SORTWEB System.**

***Index Terms*— Information Extraction, Semantics of queries, Web Search, Temporal Expressions Identification**

I. INTRODUCTION

The Web is positioned as the primary source of information in the world and the search for relevant information on the Web is considered one of the new needs of the information society. The interest of the consultation of this media is related to the effectiveness of the search engines information. The main search engines operate essentially on keywords, but this technique has limitations: thousands of pages are offered to each query, but only some contain the relevant information.

To improve the quality of obtained results, search engines must take into account the semantics of queries. The methods of information processing based on statistics are no longer sufficient to meet the needs of users to manipulate (search, translate, summarize...) information on the Web. A fact tends to be necessary: introduce "more semantic" for the search of relevant information from texts.

The extraction of specific information remains the fundamental question of our study. In this sense, it shares the concerns of researchers who have examined the texts understanding (Sabah, 2001), (Nazarenko and Poibeau, 2004), (Poibeau and Nazarenko, 1999) as those dealing today with the link between the semantic web and textual data (Berners-Lee et al., 2001), (Poibeau, 2004).

The objective of our work is to refine the search for information on the web. It is to treat the content structure and make it usable for other types of automatic processing. Indeed, when the user makes his query, he expects, generally, find precisely what he seeks, i.e. to find "the relevant information", without being overwhelmed with a volume of uncontrollable and unmanageable answers.

In the section that follows, we present some new methods which are based on analysis of the context for improving research on the Web. Then we propose our method based on two concepts: the concept of context in general (Descl és et al., 1997), (Lawrence et al. 1998) and the concept of temporal context (Faiz, 2002), (ElKhlifi and Faiz, 2010). Finally, we present the validation of our method by the SORTWEB system.

II. RELATED WORKS ON TEMPORAL INFORMATION RETRIEVAL

Nowadays, the web is operated by persons who seek information via a search engine and operate their own results. Tomorrow, the web should primarily be used by automatons that will address themselves the questions asked by people, and automatically give the best results. Thus, the web becomes a forum for exchange of information between machines, allowing access to a very large volume of information and providing the means to manage these informations. In this case, a machine can understand the volume of information available on the web and thus provide more consistent assistance to people, provided that we endow the machine with some "intelligence".

By "intelligence", we expose the fact of linking human intelligence with artificial intelligence to optimize the search of information activities on the web. The search of information involves the user in an interrogation process of the search engine. The defined query is sent to the indexes of documents. The documents whose indexes have an adequate "similarity" to the query (i.e. keywords in the query exist in the resulting documents) are considered relevant.

However, the request for information expressed by a query can be an inaccurate description of the user's needs.

In general, when the user is not satisfied with the results of its initial query, he tries to change it so as to identify its needs better. This change in the query is to be reformulated. In general, the reformulation is expressed by removing or adding words.

The results of the study by PD Bruza (Bruza and al., 2000), (Bruza and Dennis, 1997), conducted on reformulations made by users themselves have shown that reformulation is often the repetition of the initial

doi:10.4304/jetwi.4.3.234-239

request, the adding or the withdrawal of few words, changing the spelling of the request, or the use of its derivatives or abbreviations.

In this context, we can cite the system developed by HyperIndex P.D. Bruza and al. (Bruza and Dennis, 1997) (Dennis et al., 2002) relating to a technical reformulation of queries that helps the user to refine or extend the initial request by the addition, deletion or substitution of terms. The terms of reformulation, are extracted from the titles of Web pages. It is a post-interrogation reformulation: the user defines an initial query, after which the resulting titles of Web pages provided by the search system are analyzed as a lattice of terms in order to be used by the HyperIndex search engine. The user can navigate through this HyperIndex giving an overview of all possible forms of reformulation (refinement or enlargement).

Other work has been developed in this context, we can cite:

- R. W. Van Der Pol, (Van Der Pol, 2003) proposed a system to reformulation pre-interrogation based on the representation of a medical field. This field is organized into concepts linked by a certain number of binary relations (i.e. causes, treats and subclass). The complaints are built in a specification language in which users express their needs. The reformulation of requests is automatic. It takes place in two stages, the first concern the identification of concepts that pairs the need of the user, the second concerns the making up of these terms in order to formulate the request.

- A. D. Mezaour (Mezaour, 2004) proposed a method of targeted research documents. The proposed language allows the user to combine multiple criteria to characterize the pages of interest with the use of logical operators. Each criteria specified in a query can target the search for its values (keywords) on a fixed part of the structure of a page (for example, its title) or characterize a particular property of a page (example: URL). By using the logical operators conjunction and disjunction, it is possible to combine the above criteria in order to target both the type of page (html, pdf, etc.) with certain properties of the URL of a page, or characteristics of some key parts (title, body of the document). Mezaour thinks a possibility of improving its approach consists in enriching the initial request by synonyms representing the values of words for each query. According to him, the assessment of his requests passes over relevant documents that do not contain the terms of the request but equivalent synonyms.

- O. Alonso (Alonso et al., 2016) proposed a method for clustering and exploring search results based on temporal expressions within the text. They mentioned that temporal reasoning is also essential in supporting the emerging temporal information retrieval research direction (Alonso et al., 2011). In other work (Strötgen et al. 2012), they present an approach to identify top relevant temporal expressions in documents using expression, document, corpus, and query-based features. They present two relevance functions: one to calculate relevance scores for temporal expressions in general, and one with respect to a search query, which consists of a textual part, a temporal part, or both.

- In their work, E. Alfonseca et al. (Alfonseca et al., 2009) showed how query periodicities could be used to improve query suggestions, although they seem to have more limited utility for general topical categorization.

- A. Kulkarni et al. (2011), in their work, showed that Web search is strongly influenced by time. They mentioned that the relationship between documents and queries can change as people's intent changes. They have explored how queries, their associated documents, and query intents change over the course of 10 weeks by analyzing large scale query log data, a daily Web crawl, and periodic human relevance judgments. To improve their work, A. Kulkarni et al. plan to develop a search algorithm that uses the term history in a document to identify the most relevant documents.

- A. Kumar et al. (2011) proposed a language modeling approach that builds histograms encoding the probability of different temporal periods for a document. They have shown that it is possible to perform accurate temporal resolution of texts by combining evidence from both explicit temporal expressions and the implicit temporal properties of general words. Initial results indicate this language modeling approach is effective for predicting the dates of publication of short stories, which contain few explicit mentions of years.

- Zhao et al. (2012) develop a temporal reasoning system that addresses three fundamental tasks related to temporal expressions in text: extraction, normalization to time intervals and comparison. They demonstrate that their system can perform temporal reasoning by comparing normalized temporal expressions with respect to several temporal relations.

We note that, in general, manual reformulation aims at building a new query with a list of terms proposed by the system. In the case of an automatic reformulation, the system will build the new query.

However, the method of automatic reformulation, generally, does not take into account the context of the query. The standard model of search tools admits many disadvantages such as limited diversification, competence and performance. While, the establishment of research by the context is much more advantageous.

The contextual information retrieval refers to implicit or explicit knowledge regarding the intentions of the user, the user's environment and the system itself. The hypothesis of our work is that making explicit certain elements of context could improve the performance of information research systems.

The improved performance of engines is a major issue. Our study deals with a particular aspect: taking into account the temporal context. In order to improve accuracy and allow a more contextual search, we described a method based on the analysis of the temporal context of a query so as to obtain relevant event information.

III. CONTRIBUTIONS

The explosion in the volume of data and the improving of the storage capacity of databases were not accompanied by the development of analytical tools and research needed to exploit this mass of information. The realization of intelligent systems research has become an emergency.

In addition, queries for responding to requests for information from users become very complex and the extraction of the most relevant data becomes increasingly difficult when the data sources are diverse and numerous. It is imperative to consider the semantics of the data and use this semantics to improve web search. More especially as the results of a search query with a search engine returns a large number of documents which is not easy to manage and operate.

Indeed, in carrying out tests on several search engines, we found inefficient engines for queries on a date or a period of time. Therefore, we propose to develop a tool to take into account the temporal context of the query.

In this context, we propose an approach, like those aimed at improving the performance of search engines (Agichtein et al., 2001), (Glover et al., 1999, 2001) (Lawrence et al. 2001) such as the introduction of the concept of context, the analysis of web pages or the creation of specific search engines in a given field.

The objective of our work is to improve the efficiency and accuracy of event information retrieval on the Web and analyzing the temporal context for understanding the query. Therefore, the matter is to propose more precise queries semantically close to the original user's queries.

Our study consists on the one hand to reformulate queries searching for text documents having an event aspect, i.e. containing temporal markers (*i.e. during, after, since, etc.*) taking into account the temporal context of the query, and on the other hand, to obtain relevant results specifically responding to the queries.

The question that arises is how to find event information and transform collections of data into intelligible knowledge, useful and interesting in the temporal context where we are.

We found that, in general, queries seeking one or more events taking place at a given date or during a determined period do not produce the expected results. For example, *the scientific discoveries since 1940*. In this sample of query, the user wants to seek scientific discoveries since 1940 until today, not for the year 1940 only; it is then to deal with a period of time. Indeed, a standard search engine only searches on the term "1940" and not on the time period in question, from which the idea of the reformulation of the user's query, basing the search on the term introduced by the user and a combination of words synonymous with the terms of the original query. The processing of the query is mainly done at the context level. The system must be able to understand the timing of the query. Therefore, we provide it with some intelligence (to approach the human reasoning) plus a semantic analysis (for understanding the query).

Such a system is very difficult to implement for several reasons:

- The diversity of documents types on the web (file types: doc, txt, ppt, pdf, ps, etc.),
- The multitude of languages,
- The richness of languages: it is very difficult to establish a genuine process of parsing which took into account the structure of each sentence.

To do this, we will focus our work on a document type and a type of event queries containing temporal indicators (*in the month, in the year, between time and date, etc*.).

For the identification of temporal expressions, we used our method of automatic filtering of temporal information we have developed in earlier works, (Faiz and Biskri, 2002). The temporal information retrieval from the query is made by identifying temporal markers (*since, during, before, until* ...) or by the presence of explicit date in the query.

Then, for the interpretation of these terms and the need to seek event information taking place on a date or a period, we propose a time representation from the concept of interval (Allen and Ferguson, 1994). This representation is based on the start date and end date of events (*punctual or instantaneous events and durative event*).

Besides, in view of the type of queries that we will study and the temporal markers such as "before", "after" and "until", we need to express this in terms of interval. Are two types of events: punctual or instantaneous events (Evi) and durative events (EVD):

- The instantaneous event (Evi): If the beginning date is equal to the ending date of the event Deb (Evi) = End (Evi).
- The durative event (EVD): the one who takes place without interruption
  Deb (EVD) <> End (EVD).

We consider that an event E admits a start date d(E) and an end date f(E), with (d(E) <f(E)).

We ideally distinguish two types of events, those of zero duration d(E) = f (E) that are expressed, for example, by the phrase "in + (date)" example "in 2001" and whose duration is not zero, i.e. (E) <> f (E), so the interval is [d(E), f (E)] and are expressed, for example, by expression: "Between 1990 and 2001" or the phrase "since 1980. The temporal grain which we base our example is the year.

In our work, we need to represent, in the form of an interval, the temporal information contained in the query for using it as additional information to the query.

Thus, we apply the rules of interpretation to determine, in an explicit way, the time interval. Example: "If the query contains" from "+ beginning_year Then interval = [beginning_year, current_year]." So, if the document contains the event sought taking place during the time interval generated, then we consider it as relevant.

To understand the context of the query made by the user better, we also considered the extending of the query by adding words synonymous with the event in question. Example: the word "attack", we use synonyms: "attack, explosion, crime, etc ".

IV. VALIDATION OF THE PROPOSED METHOD: SORTWEB SYSTEM

We validated our work by developing SORTWEB system (System Optimization Time Queries on the Web) that improves research on the Web and through an automatic query reformulation to obtain relevant results and meet the expectations of the user. This reformulation is done by adding automatic synonymous terms sought for the event. The enrichment of the request allows for better research and results from the search terms and synonyms not only terms entered by the user.

The process is as follows:

A query such as "Event + temporal marker + date or time period" launched will be analyzed and segmented into two parts and by detecting a marker time (*in the month, in the year, since ...*).

The web search will be launched once the changes are made to the query (cf. Figure1).

- The event (containing the description of the event sought) will be transformed and reformulated referring to the basis of synonyms that allows the enrichment of the query terms in the same direction to take account of the semantics of the request.
- The part with a date or a time period which took place during the event. This part (not always the form of an explicit date, for example: last year, next year or the form of an interval, for example: during this century, since 1990) will be treated and processed under the standard form of a date or a time interval. Examples: "since 1980" will be represented by the interval [1980 2006].

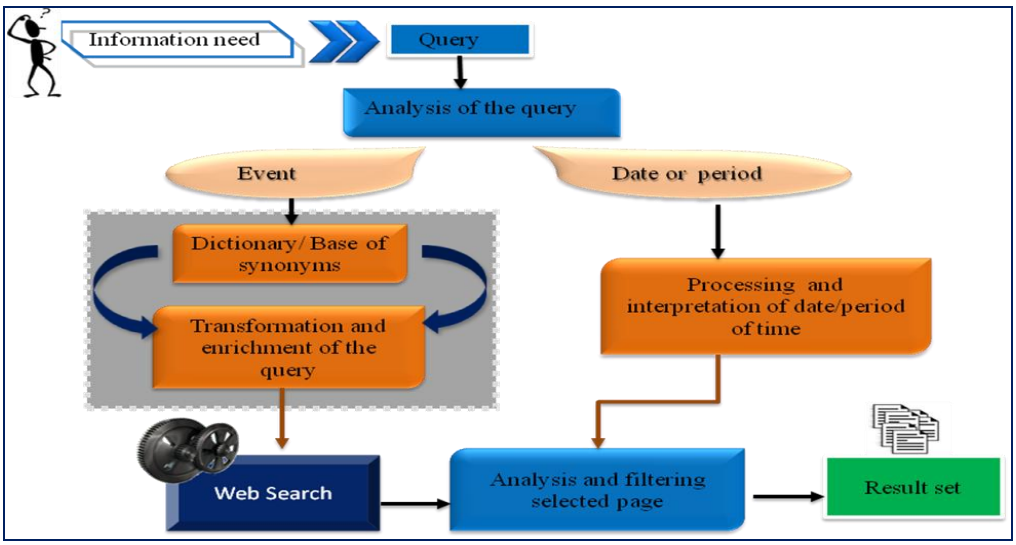

Figure 1. SORTWEB System Architecture.

This research is done using a search engine. The document is then downloaded, analyzed and then filtered by the time of the request. The filtering is to travel documents and verify results if the selected information respects the semantics of the initial request.

After the course of all documents downloaded, only the addresses of documents considered relevant are added to the page containing the results.

To test and validate our system, we launched the same requests (for example, since 2000 attacks, wars between 1990 and 2002) on several search engines such as Google or Yahoo. We ascertain that our system returns a number of documents much smaller than that proposed by them directly.

In addition, the returned results contain relevant documents issued from the search of synonyms and not from the terms of the initial request. For example: The attacks since 2000. This request is processed through our system and the research was done using the term “attack” and also the following terms “attack, explosion, crime”. The use of synonyms is very important because the user may be interested in documents containing not only the word “attack” but also containing other words in the same context.

It should be noted that the evaluation of an information retrieval system is measured by the degree of relevance of results. The problem lies in the fact that the user relevance is different from the system relevance.

In general, in a relevant document, the user may find the information he needs. We talk about user relevance when the user considers that a document meets his needs. However, the system relevance is judged through the used matching function.

To determine the relevance of obtained results, we conducted an evaluation by human experts. We found that 80% of the results were relevant.

Also, we calculated the accuracy for evaluating the quality of answers provided by the system, the results of

the tests were measured using the rate of accuracy as follows:

Accuracy = (No. relevant documents found / No documents found) = 80.6%

If our method has many advantages such as minimizing the number of such results while keeping their relevance and documents that have come from words (synonyms under the user's request) added automatically by the system, it opens up new ways of studies. One of the perspectives that we intend to achieve is the improving of the search for event information. We have to work more on the very famous events in countries where they occur such as: the event of "pilgrimage" that may be associated with "Saudi Arabia".

V. Conclusion

The new generation of search engines differs from the previous generation by the fact that these engines are increasingly incorporating new techniques other than the simple keyword search but adding other methods to improve the results of search engines, such as the introduction of the concept of context, analysis of web pages or the creation of specific search engines in a given area.

Thus, the improved performance of engines is a major issue. Our study states a particular aspect: taking into account the temporal context. In order to improve accuracy and allow a more contextual search, we described a method based on the analysis of the temporal context of a query to obtain relevant event information.

References

[1] Alfonseca, E., Ciaramita, M. and Hall, K. (2009), Gazpacho and summer rash: Lexical relationships from temporal patterns of Web search queries. In Proceedings of EMNLP 2009, 1046-1055.

[2] Agichtein E., Lawrence S. et Gravano L. (2001), Learning Search Engine Specific Query Transformations for Question Answering. Proceedings of the Tenth International World Wide Web Conference, WWW10, may 1-5.

[3] Allen J.F., Ferguson G., (1994), Actions and Events in Interval Temporal Logic. Journal Logic and Computation, vol. 4, n°5, pp.531-579.

[4] Alfonseca, E., Ciaramita, M. and Hall, K. Gazpacho and summer rash: Lexical relationships from temporal patterns of Web search queries. In Proceedings of EMNLP 2009, 1046-1055.f

[5] Alonso, O. and Gertz, M. Clustering of search results using temporal attributes. In Proceedings of SIGIR 2006, 597-598.

[6] Alonso O., Strötgen J., Baeza-Yates R. and Gertz M. (2011), Temporal information retrieval: Challenges and opportunities, TWAW 2011, Hyderabad, India, pp.1-8.

[7] Berners-Lee T., Hendler J. and Lassila O., (2001), the semantic web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American.

[8] Bruza P., McArthur R., Dennis S., (2000), Interactive internet search: keyword, director and query reformulation mechanisms compared. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, July 24-28, Athens, ACM Press, pp. 280-287.

[9] Bruza P.D. and Dennis S., (1997), Query Reformulation on Internet: Empirical Data and the Hyperindex Search Engine. Proceedings of RIAO-97, Computer Assisted Information Searching on the Internet.

[10] Dennis S., Bruza P., McArthur R., (2002), Web searching: A process-oriented experimental study of three interactive search paradigms. Journal of the American Society for Information Science an Technology, vol. 53, n°2, pp. 120-133.

[11] ElKhlifi A. and Faiz R. (2010), French-written Event Extraction based on Contextual Exploration. Proceedings of the 23th International FLAIRS 2010, AAAI Press, California, USA.

[12] Faiz R. and Biskri, I. (2002), Hybrid approach for the assistance in the events extraction in great textual data bases. Proceedings of IEEE International Conference on Systems, Man and Cybernatics (IEEE SMC 2002), Tunisia, 6-9 October, Vol. 1, pp. 615-619.

[13] Faiz R. (2002), Exev: extracting events from news reports. Actes des Journées internationales d'Analyse statistique des Données Textuelles (JADT 2002), A. Morin et P. Sébillot (Editeurs), Vol. 1, France, pp. 257-264.

[14] Faiz R. (2006), Identifying relevant sentences in news articles for event information extraction. International Journal of Computer Processing of Oriental Languages (IJCPOL), World Scientific, Vol. 19, No. 1, pp. 1-19.

[15] Glover E., Flake G., Lawrence S., Birmingham W., Giles C.L., Kruger A. and Pen-Nock D., (2001), Improving category specific web search by learning query modifications. In Symposium on Applications and the Internet (SAINT-2001), pp. 23–31.

[16] Glover E., Lawrence S., Gordon M., Birmingham W., Giles C.L., (1999), Architecture of a Metasearch Engine that Supports User Information Needs. Proceedings of the Eighth International Conference on Information and Knowledge Management, (CIKM 99), ACM, pp. 210-216.

[17] Kulkarni A., Teevan J., Svore K. M. and Dumais S. T. (2011), Understanding temporal query dynamics, wsdm WSDM'11, February 9–12, 2011, Hong Kong, China.

[18] Kumar A., Lease M., Baldridge J. (2011), Supervised language modeling for temporal resolution of texts, CIKM11, pp. 2069-2072.

[19] Lawrence S., Coetzee F., Glover E., Pennock D., Flake G., Nielsen F., Krovetz R., Kruger A., Giles C.L., (2001), Persistence of Web References in Scientific Research. IEEE Computer, Vol.34, p.26–31.

[20] Mezaour A., (2004), Recherche ciblée de documents sur le Web. Revue des Nouvelles Technologies de l'Information (RNTI), D.A. Zighed and G. Venturini (Eds.), Cépaduès-Editions, vol. 2, pp. 491-502.

[21] Nazarenko A., Poibeau T., (2004), L'évaluation des systèmes d'analyse et de compréhension de textes. In L'évaluation des systèmes de traitement de l'information, Chaudiron S. (Ed.), Paris, Lavoisier.

[22] Poibeau T., (2004), Annotation d'informations textuelles : le cas du web sémantique. Revue d'Intelligence Artificielle (RIA), vol. 18, n°1, Paris, Editions Hermès, pp. 139-157.

[23] Poibeau T., Nazarenko A., (1999), L'extraction d'information, une nouvelle conception de la compréhension de texte. Traitement Automatique des Langues (TAL), vol. 40, n°2, pp. 87-115.

[24] Sabah G., (2001), Sens et traitements automatiques des langues. Pierrel J. M. (dir.), Ingénierie des langues, Hermès.

[25] Strötgen J., Alonso O., and Gertz M. (2012). Identification of Top Relevant Temporal Expressions in Documents. In TempWeb 2012: 2nd Temporal Web Analytics Workshop (together with WWW 2012), Lyon, France.
[26] Van der Pol R.W., (2003), Dipe-D: A Tool for Knowledge-Based Query Formulation in Information Retrieval. Information Retrieval, vol. 6, n°1, pp.21-47.
[27] Zhao R., Do Q., Roth D. (2012), A Robust Shallow Temporal Reasoning System. Proceedings of the Demonstration Session at the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL 2012), pp. 29-32, Montréal, Canada.

Dr. Rim Faiz obtained his Ph.D. in Computer Science from the University of Paris-Dauphine, in France. She is currently a Professor of Computer Science at the University of Carthage, Institute of High Business Study (IHEC) at Carthage, in Tunisia. Her research interests include Artificial Intelligence, Machine Learning, Natural Language Processing, Information Retrieval, Text Mining, Web Mining and Semantic Web. She is member of scientific and organization committees of several international conferences. She has several publications in international journals and conferences (AAAI, IEEE, ACM ...). Dr. Faiz is also the responsible of the Professional Master "Electronic Commerce" and the Research Master "Business Intelligence applied to the Management" at IHEC of Carthage.