Contextual hybrid-based recommendation of PUBMED articles

Type of article: Technical note.

Mohammed Ilyas Tchenar[1], Youssouf Rahali[1], Abdeldjalil Khelassi[1,2]

[1]*Knowledge and information Engineering Research Team, University of Tlemcen Algeria.*

[2]*Informatic Department, Sciences Faculty, University of Tlemcen Algeria.*

Khelassi.a@gmail.com

Abstract

Background: The amount of information at the medical field has been growing day by day. Also new medical articles about the preoccupied disease are published each day and the updated information is very required for physicians. Consequently, the appropriate information in the appropriate moment is very suitable for physicians, and this is the challenge contributed in this research work.

Methods: Our goal is to recommend documents deemed relevant to doctors, moreover, regarding the context of using of an Electronic Medical Records EMR. The principle is to extract the context of this usage: illness, Age ..., searching on the contents of documents and taking into account the rate of vote documents. For experiment and evaluation, we have used 100 articles randomly selected from PUBMED about cardiology. In addition, we have developed a system that extracts the context of EMR at the moment of exploration. The extracted context is used with users rating by the recommender system to select and rank the recommended articles for physicians in the same moment of use.

Results: The first result of this research work is the smart interaction between users and the software system by introducing the context of use. In addition, another important result is the reuse of user's appreciation for additional dynamicity and intelligibility.

Conclusion: The developed system offers the physician an appropriate recommendation of pertinent PUBMED articles. The developed system augments the relevancy of the recommendation by analyzing the contents of the articles and introducing a collaborative method.

Keywords: *Information filtering, context, recommendation, hybrid-based recommendation, medical record system, cardiology.*

1. Introduction

Nowadays, Internet users are struggling to find the data and information they seek since of the information overload on the web, to solve the problem the systems of recommendations have appeared. The goals are to minimize the user's time spent on searching and suggest relevant resources that they would not have consulted.

Several approaches are developed for recommending resources [1]: 1) The collaborative approach based on user's opinion and similarity between users, 2) the content approach that makes recommendations by comparing the content of resources with the user's sensitivity, 3) the hybrid approach that combines the previous approaches.

In addition, the context-aware is an important opportunity for developers to integrate an adapted behavior in the intelligent systems. It permits to understand the user to offer personalized services, which appear by consequence, smart [2].

This research work is a part of the project “RAMHeR: Reuse and Mining Health 2.0 Resources” [3]. The problem we are dealing with is related to linking the recommendations of selected PUBMED documents with Electronic Medical Record EMR users, by exploiting hybrid information filtering approach, which applies collaborative vote, collected from other users and the transparent extraction of context. The content of the documents is also used for information retrieval.

In this work, we adopt a contextualized recommendation approach, in which the query is transparently extracted from the context of application use. The extracted information from the context of EMR uses as: its disease, its age, its sex.. Etc. Moreover, we have also applied the hybrid recommendation approach, in which the content of documents is analyzed by the “term of frequency/inverse document frequency” tf/idf technic [4]. In addition, it is collaborative by recommending documents that have higher rate of votes.

2. Methods and results

The main objective of our work is to produce an intelligent recommendation of PUBMED documents. The recommendation is contextual which is based on patient information extracted from an EMR figure 7. It is also hybrid based on indexing the documents contents to find the most important words by using tf/idf technic figure 4, which calculates the similarity between document and the query extracted from context. In this technic, we have applied the cosine similarity measure enriched by the voting rate of the documents. For demonstration in Figure 8 the selection of pertinent documents has been realized in a corpus collected from PUBMED which is relative to the cardiology domain. The corpus contains 100 documents and the selection was randomly by recording all results of the query “CARDIOLOGY”. In addition, the vote was collected from some students in several uses.

2.1 The conceptual models

All diagrams are written in Unified Modeling Language UML. In figure 1, we describe the uses cases diagram. The utilization of our system is only authorized by the physician to protect the privacy of patient's medical record. The second application of the system refers to the documents recommended by the system.

These two applications are described also in figure 2 and 3 by a sequences diagram, which presents the different scenarios of actions carried out by the actors of the system. Moreover, in figure 2, (Authentication and visualization of medical record), the doctor enters the pseudo and password details, the system checks this information and returns the answer, once it is done the doctor looks for the record of a patient's follow-up. In figure 3 (Extraction of context and recommendation), the doctor begins by entering a new patient by completing the follow-up form (name, surname, illness, age, etc.). Following that, the system records this new follow-up record and extracts the patient's context (sickness, age, sex). After that, the system compares, transparently, the constructed query from the context with the documents for information retrieval. Consequently, the system recommends the documents that it finds appropriate to the case.

2.2 Content recommendation

a. Content indexing process

Indexing involves parsing each document in the collection to create a set of terms. These terms will be easily exploited by the system during the subsequent search process. Indexing phase is applied to create a representation of documents in the system. Ordinarily, the objective is to find the best important concepts of the document, which will form the descriptor of the document. Moreover, indexing can be:

-Manual: A specialist in the text-field analyzes each document, to ensure a better relevance for the answers provided by the IRS, but the time necessary for its realization is right important.

-Automatic: the indexing process is entirely computerized; it groups together a set of automated treatments on a document. We can distinguish: automatic extraction of document words, elimination of empty words, lemmatization (radicalization or normalization), identification of groups of words, weighting of words and consequently the creation of the index.

-Semi-automatic: the final choice belongs to the specialist or the documentarist, who often intervenes to choose other significant terms. The indexers use a thesaurus or a terminological database, which is an organized list of descriptors (terms) that follows their own terminological rules and are linked by semantic relations. The choice and the interest of a method over others dependence on a certain number of parameters, the most decisive of which is the volume of the collections.

b. Lexical analysis

Lexical analysis is the process of converting the text of a document into a set of terms. A term is a lexical unit or a radical. Moreover, lexical analysis ensures the recognition spaces of words' separation, numbers, punctuations, etc.

- Weighting of terms: The weighting of terms is used to measure the importance of a term in a document. This importance computing is often based on statistical (or sometimes linguistic) considerations and interpretations. The goal is to find the terms that best represent the content of the document. If we list the completely different words of any text in descending order of frequency, we note that the frequency of a word is inversely proportional to its rank in the list. The law of Zipf formally states this statement:

$$\text{Rank} * \text{frequency} = \text{constant} \quad (1)$$

The relation between the frequency and the rank of the terms permits to select the terms representative of a document: the very high frequency terms are eliminated since they are not representative of the document (for example, the word "tools"), and the very low frequency terms (which eliminates keystrokes).

Ordinarily, weighting techniques are based on the factors tf and idf, which combine the local and global weightings of a term:

tf (Term Frequency): this measurement is proportional to the frequency of the term in the document (local weighting). It can be used as it is or in several variations (log (tf), presence / absence, ...).

idf (Inverse of Document Frequency): This factor measures the importance of a term throughout the collection (global weighting). A term that often appears in the documentary database should not have the same impact as a less frequent term. It is generally expressed as follows:

$$\text{idf} = \log (N / \text{df}), \quad (2)$$

Where:
· df is the number of documents containing the term.

· N is the total number of documents in the database.

Measure tf * idf: By combining the two previous technics, it gives a good approximation of the importance of the term in the document, particularly in a corpus of documents of homogeneous size. However, it does not consider an important aspect of the document: its length. In general, the longer documents tend to use the same terms repeatedly, or to use additional terms to describe a topic. Therefore, the frequencies of the terms in the documents will be higher, and the similarities to the query will also be greater. To overcome this disadvantage, it is possible to integrate the size of the documents into the weighting formula: this is called a normalization factor.

-Index creation: In order to respond more quickly to a request, special storage structures are required to store the selected information during the indexing process. The most common means of storage are: inverted files, suffix arrays, and signature files. Inverse files are currently the best choice for several applications. The inverse files consist of two main elements:

- The vocabulary, which is the set of all the different words of the text;

- Posting: for each word, it is a list of all the positions in the text for which the word appears.

Moreover, figure 4 shows an example of vocabulary and occurrences.

c. matching: Document-query

The comparison between the document and the request amounts to calculating a score, supposed to represent the relevance of the document to the request. This value is calculated from a similarity function or probability noted RSV (Q, d) (Retrieval Status Value), where Q is a query and d a document. This measure considers the weight of the terms in the documents, determination is based on statistical and probabilistic analyzes. The matching function is closely related to the indexing and weighting operations of the query terms and the corpus documents. In general, the document-request matching and the indexing model permits to characterize and identify an information retrieval model. The similarity function, then allows the documents returned to the user to be ordered. The quality of this scheduling is paramount. In our system, the query is the context and the measure of similarity used is the cosine.

2.3 Collaborative Recommendation:

Our system gives physicians the opportunity to vote on the relevancy and the pertinence of documents, see figure8. Via the system, the votes are collected from physicians for computing the average of the votes. The system in the future uses recommends the voted documents to the other doctors in relation with the rate of votes and the average. Since all the doctors are cardiologists then one does not calculate the measure of similarity between them.

2.4 Human-machine interface description

Figure 5 to 8 demonstrate the interaction between physicians and the developed system as a web application. Figure 5 and 6 are the first pages of the web site, it allows physicians to register to be authenticated by the system in their future use. The interface in figure 6 facilitates the authentication of registered users. This authentication ensures the privacy of the physicians and the patients allocated to him in the EMR. By authentication the physician is allowed to edit and consult the electronic records of his patients. Figure 7 shows the possibility to realize a search based on the patients records. Figure 8 presents the results of our system, in which after showing the electronic medical record of the selected patient, a ranked list of PUBMED articles is presented ordered by the mean measure of content pertinence and the physicians' votes.

3. Discussion

In this technical note, an original web application is presented in which we have introduced a new contextualized recommendation of appropriate documents in the phase of follow-up recording or visualizing. This contribution is not just for ensuring collaboration between physicians who uses this application, but also to ensure a preventive learning of newest research innovation and finding in the right moments.

Ordinarily, the application at first shows the patient's follow-up sheet, moreover, at the appropriate moment of usage, several EMR services are offered by the application. The visualization of patient's follow-up sheet transparently start the second task, which is the extraction of context of use. After that, the system edits the query to be compared to the documents of the corpus. By applying information retrieval technique, the system displays the relevant documents to the patient case. Once the doctor clicks on the document link, he will be directed directly to the PUBMED document. Following that, he could introduce his feedback by a vote on that document only once.

Consequently, the success of the application is presented in figure 8 in which a significant example was explored. In this example, the list of recommended articles is appropriate to the illness and therapy of the patient, also the doctors' votes appear behind each article title and link to PUBMED. Moreover, the users could recommend articles for others, which ensure some curative feedbacks by physicians.

For demonstration and tests, we have collected a corpus of document containing 100 articles from PUBMED in cardiology. In addition, the context is extracted from EMR by a self-developed algorithm in the exploration phase . The extracted context is used with users rating by the recommender system to select and rank the recommended articles for physicians in the same moment of use. This trend resolves the problem of the cold phase in recommending systems, where the documents contents are used in the absence of votes.

4. Conclusion

Nevertheless, the system presents a smart approach to offer the pertinent information at the right moments; several inclusions are scheduled for the future versions of this system for additional improvements. These inclusions like information retrieval optimization and improvement to improve the precision and transparency of the system. By applying the existing semantic methods, moreover, by extending the elements of context. For this last, the social information can be scheduled too.

5. Acknowledgment

We would like to thank all students who contribute in this project.

6. Authors' biography

Dr. Eng Abdeldjalil Khelassi (Corresponding author) is an associate professor of Informatics at Abou Bekr Belkaid University of Tlemcen. Head of Knowledge and Information Engineering Research Team. Associate editor at Electronic Physician.

Mr.Mohammed Ilyas TCHENAR received the Master degree in Knowledge and information systems engineering from faculty of Science, Abou Bekr Belkaid university of Tlemcen, Algeria, in 2014. He is currently pursuing the PhD degree in Spatial Information Engineering and remote sensing at the School of Computer Science and Engineering, Beihang University, Beijing. His research interests include pattern recognition, remote sensing applications, change detection, computer vision, and image processing.

Mr. Youssouf RAHALI is a computer engineer, he graduates from faculty of Science, Abou Bekr Belkaid university of Tlemcen, Algeria, in 2014 with a master's degree in knowledge and information systems engineering. He works today as a web developer.

7. Conflicts of interest

The system presented in this article was funded as master thesis of Mohammed Ilyas Tchenar and Youssouf Rahali advised by Dr Abdeldjalil KHELASSI at Computer sciences department, Sciences Faculty, Abou Bekr Belkaid University of Tlemcen.

8. References

1. Khelassi, A. (2016, January). An augmented pragmatics by explanation-aware and recommendation-aware in the context of decision support. In *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)* (p. 79). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).
2. Baldauf, M., Dustdar, S., & Rosenberg, F. (2007). A survey on context-aware systems. *International Journal of Ad Hoc and Ubiquitous Computing*, *2*(4), 263-277.
3. Khelassi, A. (2015). RAMHeR: Reuse And Mining Health2. 0 Resources. *Electronic Physician*, *7*(1), 969.
4. Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*.

5. Figures

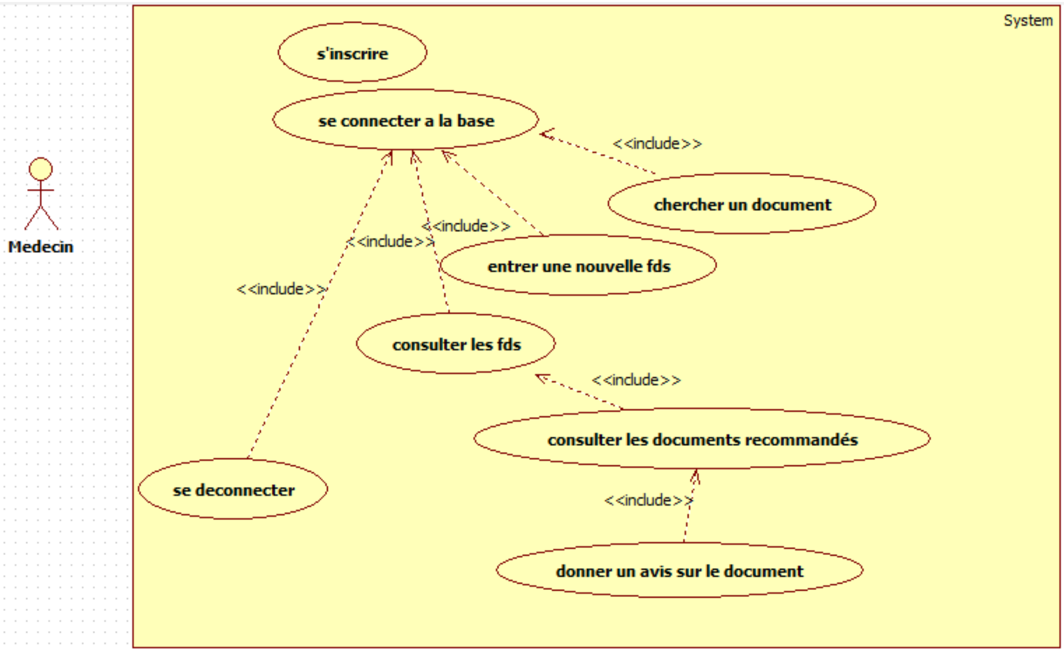


Figure .1: Use Case Diagram

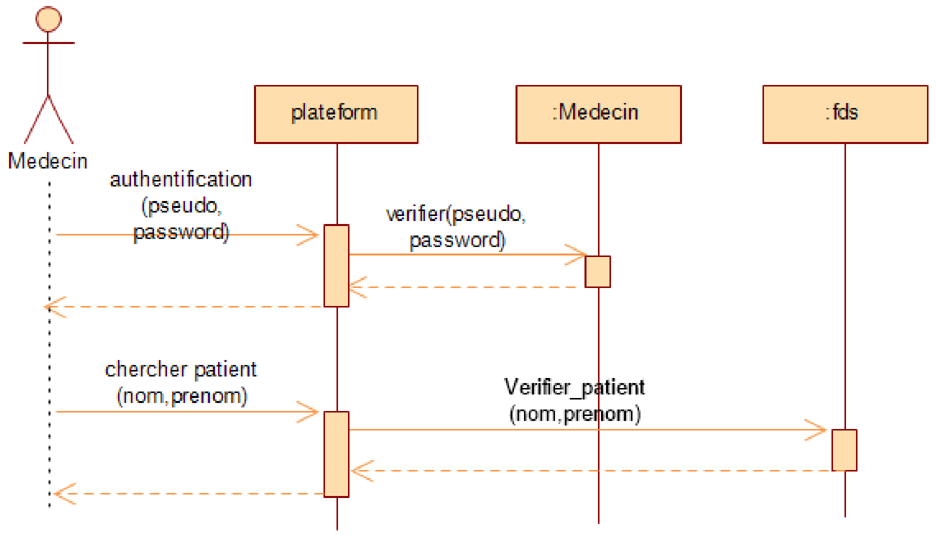

Medecin
plateform
:Medecin
:fds
authentification (pseudo, password)
verifier(pseudo, password)
chercher patient (nom,prenom)
Verifier_patient (nom,prenom)

Figure .2 sequence diagram for authentication and search

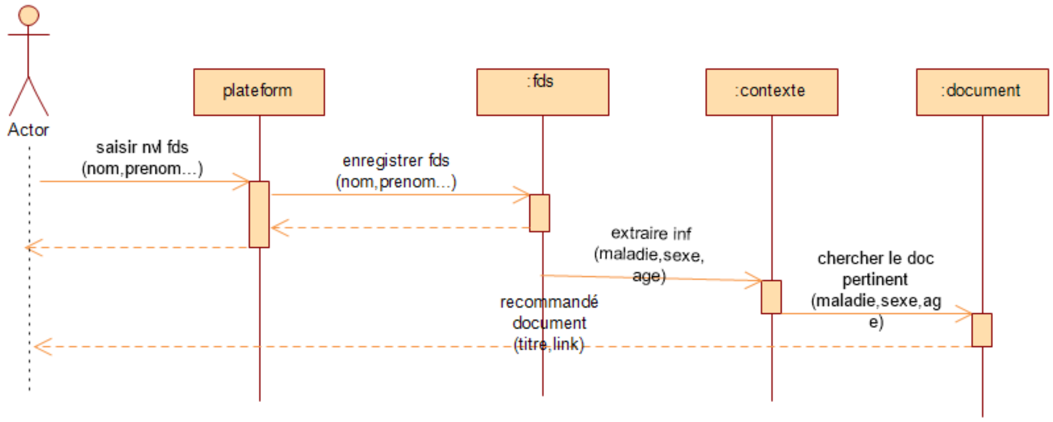
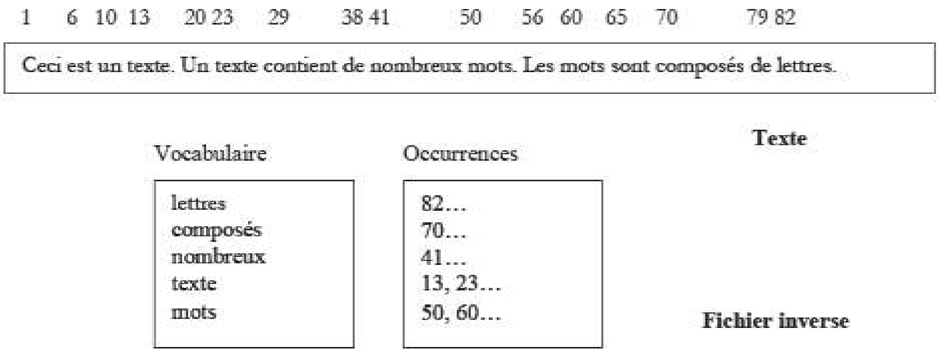

Actor
plateform
:fds
:contexte
:document
saisir nvl fds (nom,prenom...)
enregistrer fds (nom,prenom...)
extraire inf (maladie,sexe, age)
chercher le doc pertinent (maladie,sexe,age)
recommandé document (titre,link)

Figure .3 Sequence diagram of Context Extraction and Document Recommendation



Ceci est un texte. Un texte contient de nombreux mots. Les mots sont composés de lettres.

Vocabulaire	Occurrences
lettres	82…
composés	70…
nombreux	41…
texte	13, 23…
mots	50, 60…

Figure .4 A simple text and the corresponding reverse file

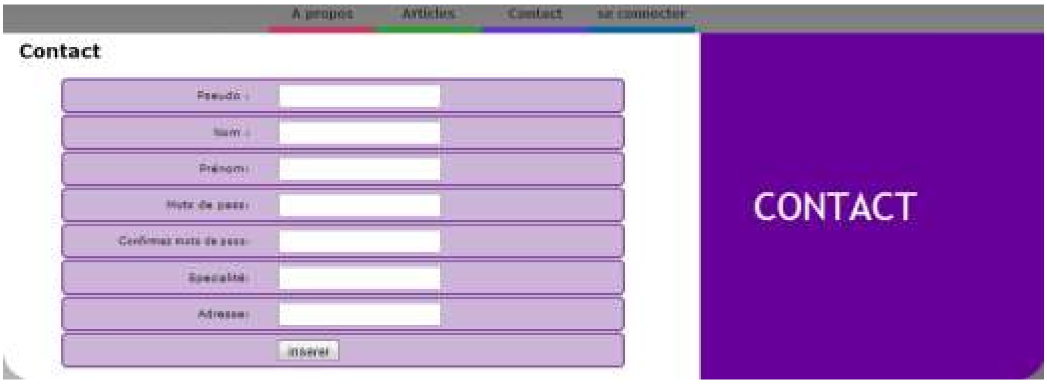


Figure .5 Welcome and registration web page

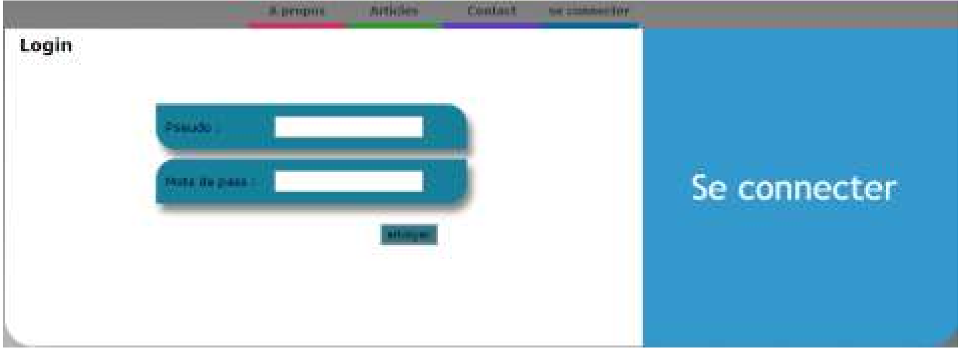
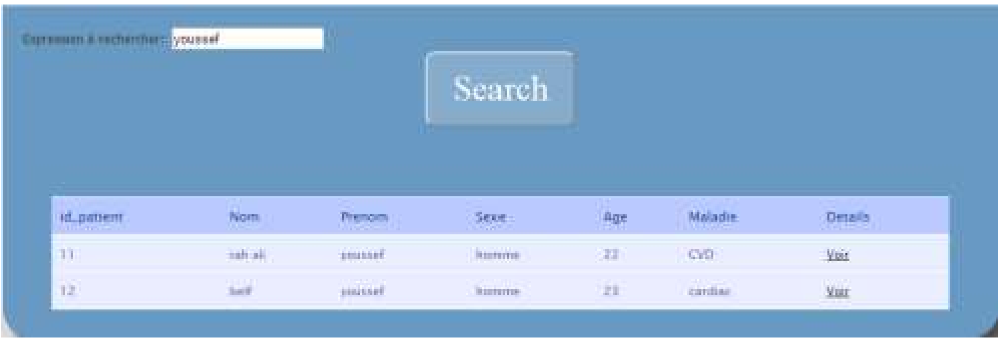


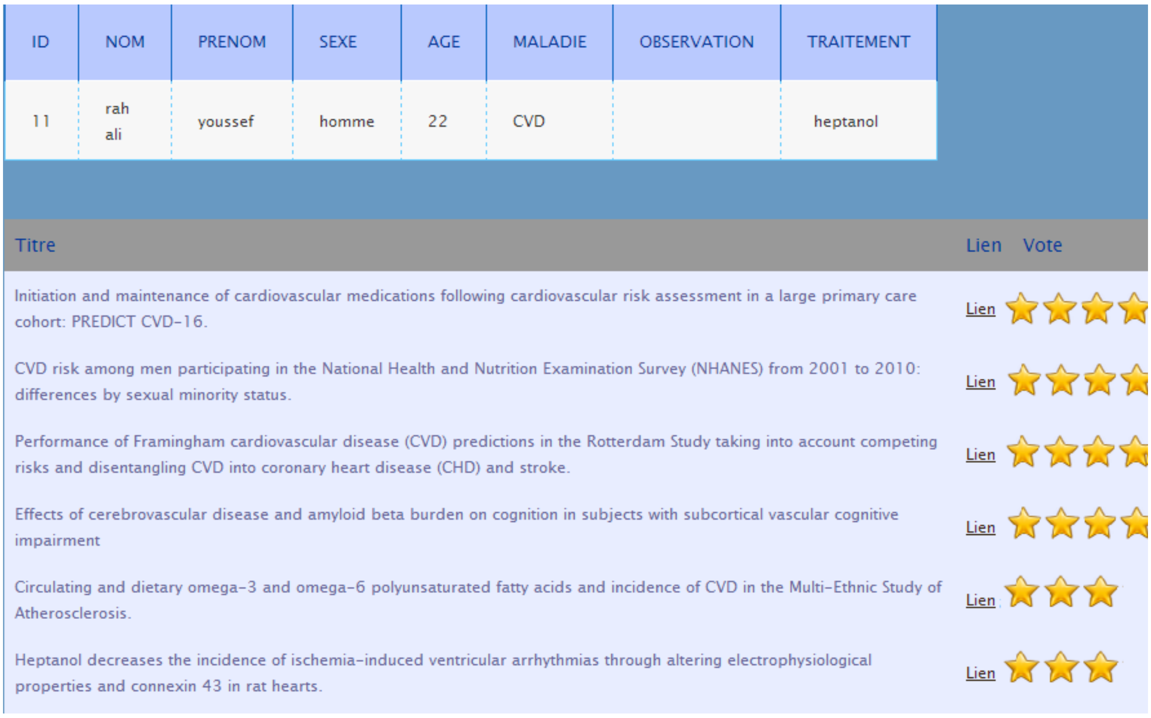
Figure 6 Connection web page



Expression à rechercher : youssef

id_patient	Nom	Prenom	Sexe	Age	Maladie	Details
11	sah ali	youssef	homme	22	CVD	Voir
12	seif	youssef	homme	23	cardiac	Voir

Figure .7 searching page



ID	NOM	PRENOM	SEXE	AGE	MALADIE	OBSERVATION	TRAITEMENT
11	rah ali	youssef	homme	22	CVD		heptanol

Titre	Lien	Vote
Initiation and maintenance of cardiovascular medications following cardiovascular risk assessment in a large primary care cohort: PREDICT CVD-16.	Lien	★★★★
CVD risk among men participating in the National Health and Nutrition Examination Survey (NHANES) from 2001 to 2010: differences by sexual minority status.	Lien	★★★★
Performance of Framingham cardiovascular disease (CVD) predictions in the Rotterdam Study taking into account competing risks and disentangling CVD into coronary heart disease (CHD) and stroke.	Lien	★★★★
Effects of cerebrovascular disease and amyloid beta burden on cognition in subjects with subcortical vascular cognitive impairment	Lien	★★★★
Circulating and dietary omega-3 and omega-6 polyunsaturated fatty acids and incidence of CVD in the Multi-Ethnic Study of Atherosclerosis.	Lien	★★★
Heptanol decreases the incidence of ischemia-induced ventricular arrhythmias through altering electrophysiological properties and connexin 43 in rat hearts.	Lien	★★★

Figure .8 Recommendations and voting page