InfoRoute: the CISMeF Context-specific Search Algorithm

Tayeb Merabti[a], Romain Lelong[a], Stefan Darmoni[a,b]

[a] *CISMeF & TIBS, LITIS EA 4108, Rouen University Hospital, Rouen, France*
[b] *INSERM, U1142, LIMICS, F-75006, Paris, France*

Abstract

Objective*. The aim of this paper was to present a practical InfoRoute algorithm and applications developed by CISMeF to perform a contextual information retrieval across multiple medical websites in different health domains.* ***Methods.*** *The algorithm was developed to treat multiple types of queries: natural, Boolean and advanced. The algorithm also generates multiple types of queries: Boolean query, PubMed query or Advanced query. Each query can be extended via an inter alignments relationship from UMLS and HeTOP portal.* ***Results.*** *A web service and two web applications have been developed based on the InfoRoute algorithm to generate links-query across multiple websites, i.e.: "PubMed" or "ClinicalTrials.org".* ***Conclusion.*** *The InfoRoute algorithm is a useful tool to perform contextual information retrieval across multiple medical websites in both English and French.*

Keywords:

Algorithms, information storage and retrieval, medical informatics applications.

Introduction

The Internet and in particular the Web has become an extensive health information and knowledge repository. As described in the survey published by Podichetty et al. [1], a majority (72%) of physicians affirmed that they used the Internet on a regular basis for medical research and 51% of them confirmed that the Internet influenced their healthcare practice. However, as mentioned by De Leo et al. [2] the vast majority of physicians (92%) indicated that they access a target site rather than utilize a search engine to gather medical information. A comparison study by Freeman et al. [3] concluded that PubMed appears to be more specific than Google Scholar; the same conclusion was reported by Anders et al.[4] "PubMed is more practical to conduct efficient and valid searches on clinical topics than Google Scholar".

With the explosion of available health information and the number of biomedical websites in different medical fields including: literature databases, clinical trials databases, and drug databases; knowledge and information retrieval (IR) have become more complex and more time consuming. In the medical field some problems have obstructed the IR [5],[6] including the following:

- **Inadequate expression of information needs:** many information needs cannot be appropriatley expressed: query syntax, medical terminologies or languages.
- **Lack of the relevant information:** Due to the inadequate terms used to search information, as medical databases and IR systems are indexed according to specific health terminologies which are different: lexically (different terms referred to the same medical concept) and structurally (different relationships between same concepts).

The aim of this work is to propose an algorithm named "InfoRoute" to perform a contextual and cross-lingual information retrieval across multiple websites in different health fields. This algorithm is devoted to French-speaking health professionals, who can read English but who prefer to perform queries in their native language. InfoRoute builds and generates various IR query on websites such as: PubMed, ClinicalTrials.org. Besides the algorithm, a web service and multiple web applications have been developed to help physcians and students to express the adequate IR query including medical terms in English and French based on more than 56 health terminologies.

Materials

The UMLS Metathesaurus

The UMLS Metathesaurus[7], developed by the US National Library of Medicine (NLM®) integrates 2,930,638 concepts in its 2014 release from 168 biomedical vocabularies. In the "MRCONSO"[1] table, which lists all UMLS concepts, only four terminologies, of all UMLS terminologies, are included with their French version in the UMLS Metathesaurus: the MeSH thesaurus, the World Health Organization Adverse Reaction Terminology (WHO-ART), the WHO International Classification of Primary Care (ICPC2), and the Medical Dictionary of Regulatory Activities (MedDRA). However, five (5) Biomedical Terminologies and Ontologies (BMTO) that have an existing official French version, are included in the UMLS but without their French version: the International Statistical Classification of Diseases (ICD10), the Systematized NOmenclature of MEDicine (SNOMED Int), Logical Observation Identifiers Names and Codes (LOINC), the International Classification of Functioning, Disability and Health (WHO-ICF) for handicap and the International Classification for Nursing Practice (ICNP). Furthermore, the CISMeF team has partially translated the BMTO included in the UMLS only in English: 24,563 synonyms and 689 ambiguous acronyms of the MeSH Descriptors, 163 synonyms of the MeSH Qualifiers, 20,887 MeSH Supplementary Concepts, and, 847 MEDLINEPlus terms and 12,700 FMA terms.

[1] The name of the concepts table in the UMLS metathesaurus

The Health Multiple Terminologies and Ontologies Portal (HeTOP)

A generic meta-model was designed in order to fit all 56 terminologies into one global structure.

The HeTOP[8] is connected to this meta-model to search concepts from all the health terminologies available in French (or in English and translated into French) included in this portal and, to browse it dynamically. This allows to:

a) Manual or automatic indexing of resources for the catalogue;
b) Retrieval of resources;
c) Teaching or performing audits in terminology management.

Some terminologies and classifications are included in the UMLS Metathesaurus (N=17) but the majority is not (N=39), i.e. the Human Rare Diseases Ontology (HRDO). Currently, HeTOP integrates 1,743,772 concepts in English, 1,031,230 in French, and 9,255,438 relationships.

The HeTOP portal integrates multiple terminological relationships and they can be classified as:

1. Intra Terminological relationships: terms linked with this type of relationships are in the same terminologies. For example, the MeSH term "suicide" is linked according to the Intra MeSH relationships **"See** also**"** with the MeSH term "suicide, attempted".
2. **Inter Terminological relationships:** these relationships link terms from different terminologies. In HeTOP, four (4) main inter-relationships were integrated and/or created (see *Table 1* for examples):
   a. UMLS alignement ($UMLS_{alignment}$) :
      This conceptual relationship described previously by Merabti et al. [9]. Two terms in HeTOP are linked under this relationship if they share the same UMLS concept in the Metathesaurus.
   b. Manual inter-alignment ($CISMeF_{manual}$):
      This relationship is added manually in HeTOP by CISMeF physicians or terminological specialist to link between two similar terms described in the same medical concept.
   c. Exact inter-alignment ($CISMeF_{exact}$):
      This relationship is obtained automatically using the lexical approach previously described by Merabti et al.[9]. Two terms are linked using $CISMeF_{exact}$ if they are lexically similar at the preferred term or at the synonym level, in French or English.
   d. Exact supervised inter-alignment ($CISMeF_{supervised}$):
      There are several $CISMeF_{exact}$ relationships validated by CISMeF physicians and tagged as $CISMeF_{supervised}$. In terms of efficiency this relationship is equivalent to $CISMeF_{manual}$ relationship.

Methods

Websites categorization

A total of twelve (12) catagories were created to classify multiple medical websites (see Table 2). The selection of websites was performed by librarians according to the needs of health profesionals, students and patients. Moreover, some websites were considered as the most important: e.g. PubMed, US Clinical Trials and EU Clinical Trials. Therefore, specific functions for these websites were developed. For example, in the following sections we described a PubMed query which is a specific quering on PubMed. The websites selected are important since it conditions the input and ouput of the algorithm developed.

Table 1 –examples from each 4 inters relationsships integrated and their number in HeTOP (2014 released)

	Source Term (Terminology)	**Target Term (Terminology)**	**Number of relations in HeTOP**
$UMLS_{alignment}$	*Myocardial Infarction (MeSH)*	*Myocardial infarction, NOS (SNOMED Int)*	*644,982*
$CISMeF_{manual}$	*Riedel thyroiditis(HRDO)*	*Riedel's thyroiditis (MedDRA)*	*41,673*
$CISMeF_{exact}$	*appetite stimulants (ATC)*	Appetite stimulated (WHOART)	*653,709*
$CISMeF_{Supervised}$	*gonadotropin-releasing hormone (MeSH)*	*Luteotropin-releasing factor (FMA)*	*251,995*

InfoRoute Algorithm

InfoRoute is defined as an algorithm that automatically generates multiple search queries across many medical websites and exploiting the entire range of French and English Medical terminologies included in the HeTOP and some terminologies from the UMLS Metathesaurus.

The input

The InfoRoute algorithm takes in input three (3) kinds of queries:

1. **Natural Input Query (NI_{Query}):** this type of query assumes that the input text is composed of multiple terms in English and French in the natural language without using Boolean or restricted operators.
2. **Boolean Input Query (BI_{Query}):** query is based on the CISMeF Boolean query syntax shown in an equation 1. Where:
   a. **RO:** corresponds to the restricted operators (ROs)[2] as defined by the CISMeF such as "mr" for reserved words, "ti" for title or "tc" as all fields. These ROs can be useful to accurately translate the query to PubMed.
   b. **TERMINO:** the terminologies assigned to each term.

In addition to these ROs, the following terms can be connected

(Term(.RO([TERMINO])*)$^{+}$((AND|OR|NOT)(Term(.RO([TERMINO])*)$^{+}$)* (1)

using Boolean operators: AND, OR and NOT.

For example, the query **"asthma.mr[TER_MSH] AND child.tc"** corresponds to the query "search resources indexed

[2] Restricted operators used in the CISMeF search engine: http://doccismef.chu-rouen.fr/aides/aidedcacronyme.html [Nov 2014]

by asthma as a reserved MeSH term and with the term child in all fields (title, abstract, etc.).

3. **Advanced Input Query (AI_{Query}):** this type of query is composed of at least one natural or Boolean query and some options such as: age, country, gender. These options are used to translate the query to clinical trial websites: ClinicalTrials, Clinical Trials Register.

In addition to the query, the algorithm is a terminology depending, therefore, according to the terminology assigned in an input, and the results change whether query terms are included or not.

The Mutli-Terminological Automatic Indexing Query (MTAIQ)

This part of algorithm is very important, since in this query stage the input is indexed to extract the most similar medical terms in the query. The MTAIQ is a multi-terminological-bilingual indexer [9]. It uses the HeTOP databases and natural language processing tools [9] to analyze and normalize the query. In the case of the natural query, MTAIQ maps the query to the most similar term(s) depending on terminologies assigned in the input. For example, if the query is **"childhood asthma"** then MTAIQ detects "asthma in children" term, since the "childhood asthma" is a synonym of the MEDLINEPlus term "asthma in children". Nevertheless, if only the MeSH terminology is assigned then the most similar term detected will be "**asthma**", since "childhood asthma" and childhood are not MeSH terms.

The query Translation

In order to query all the websites, the InfoRoute algorithm performed multiple query translation from the one of the three queries types to at least four (4) possible kinds of queries:

1. **Boolean Output Query (BO_{Query}):** the query generated is composed by terms indexed using the MTAIQ and separated with Boolean operators. For example, if the query is **"furlong syndrome"** and the terminology selected is MeSH then the query generated will be "(**Furlong syndrome) OR (marfanoid disorder with craniosynostosis, type 2".** In this example the entry and the input languages were in English but it is possible to translate the result into the corresponding French Boolean query: **(syndrome de furlong) OR (furlong) OR (craniosynostose marfanoide)**".

2. **Extended Boolean Output Query (EBO_{Query}):** the Boolean query generated can be extended using the inter terminology relationships: $UMLS_{alignement}$, $CISMeF_{manual}$ and $CISMeF_{supervised}$ relationships described above. The aim was to obtain a new query with additional synonyms not included in the original user query. In this case, synonyms are all terms preferred or synonyms related to the query terms in the same language and according to the medical terminologies assigned.

3. **PubMed Query ($PubMed_{Query}$):**

Most of the research efforts were concentrated on PubMed queries because PubMed is of utmost importance for health professionals, students and patients. Some improvements have already been performed by CISMeF on PubMed query (vs. the PubMed query by default) since 2009 [11],[12]. The first study was testing the added value of MeSH synonyms (MeSH Entry Terms) [11]] and the second study was testing the added value of UMLS synonyms (same CUI) [12]. The PubMed query is a Boolean query combined MeSH terms (preferred and synonyms) and search options from PubMed: [MH]: Main Headings, [TI]: Title or [TW] for text words. This query was generated from natural or Boolean query. In addition to the simple PubMed query, in CISMeF two additional PubMed extended queries and one PubMed manually build query were developed:

a. **PubMed UMLS-Extended Query ($PubMedUmlsE_{Query}$):** As the EBO_{Query} query and as described in [12]. The PubMed query was extended by adding synonyms from UMLS which represent terms which are in relation with the MeSH terms in the query and translated into English.

b. **PubMed Inter Extended Query ($PubMedInterExt_{Query}$):** Like the $PubMedUmlsE_{Query}$, the PubMed query was extended by adding synonyms which are in related to the MeSH terms according to two (2) inter relationships: $CISMeF_{manual}$ and $CISMeF_{supervised}$. Furthermore, for some BMTO included in HeTOP a PubMed generator query was developed based on these relationships. This generator extracted only the MeSH terms (MeSH descriptor and MeSH Supplementary concepts) for "no MeSH terms" which were related to it and build automatically a PubMed query using these MeSH terms. For example, for the HRDO term "Marfan syndrome" will be extended to the MeSH term "marfan syndrome" and all its synonyms.

c. **PubMed Manual Query ($PubMedManual_{Query}$):** In some cases and because the automatic generated query was not effective (for the drug terms as an example), a manually PubMed query for each term was created in order to be more accurate. In a previous study[13], the "ATCtoPubMed" application to access PubMed via any Anatomical Therapeutic Chemical Classification (ATC) code was described. For each ATC code, a predefined query was created and could be entered on PubMed.

4. **Advanced Clinical Query ($Advanced_{Query}$):** This query is based on the AI_{Query} since it generates a "Boolean" query combined with the AI_{Query} Options selected. Currently, options have been selected based on two clinical trials websites: http://clinicaltrials.gov (US clinical trials) and http://www.clinicaltrialsregister.eu (EU clinical trials). Moreover, one name options were used as input and the $Advanced_{Query}$ mapped options in input to the adequate name option in the query search for each clinical websites. For example, the option "SEXE=female" will be mapped to: "gender=female-only" in EU clinical trials and to "gndr=Female" in the US clinical trials.

Table 2– Principals Websites classified by categories and language.

Categories	Website	
	English	**French**
Clinical Trials	1. ClinicalTrials: http://clinicaltrials.gov 2. Clinical Trials Register: http://www.clinicaltrialsregister.eu	1. ANSM: http://ansm.sante.fr
Drugs	1. MedlinePlus: http://www.nlm.nih.gov/medlineplus/	1. Portail D'information sur les medicaments: http://doccismef.chu-

	2. Drug Information Portal: http://druginfo.nlm.nih.gov/drugportal	rouen.fr/servlets/PIM http://www.has-sante.fr/
Information For Patients	1. NIH Senior Health: http://nihseniorhealth.gov/ 2. MedlinePlus: http://www.nlm.nih.gov/medlineplus/	1. CISMeF Patient: http://doccismef.chu-rouen.fr/dc/#env=pa
Public Health		Banque de données en santé publique http://www.bdsp.ehesp.fr/
Rare Diseases	1. Gentics Home reference: http://ghr.nlm.nih.gov/ 2. NCBI OMIM: http://www.ncbi.nlm.nih.gov/omim?	1. Orphanet: http://www.orpha.net
Search Engine	1. PubMed: http://www.ncbi.nlm.nih.gov/pubmed 2. NLM Gateway: http://gateway.nlm.nih.gov/gw/Cmd/home.jsp	1. CISMeF: http://doccismef.chu-rouen.fr/dc/
Students	1. PubMed: http://www.ncbi.nlm.nih.gov/pubmed	1. DocUMVF: http://doccismef.chu-rouen.fr/dc/#env=umvf

Table 3– Examples of queries types generated by InfoRoute Algorithm.

Query Search Type	Query	Query Terminology	InfoRoute query
BO $_{Query}$	abnormal platelets	*MedDRA*	(abnormal platelets) OR (platelets abnormal) OR (thrombocytes abnormal (nos))
EBO $_{Query}$	abnormal platelets	*MedDRA*	(abnormal platelets) OR (platelets abnormal) OR (thrombocytes abnormal (nos)) OR (platelet abnormalities) OR (abnormal platelet) OR (atypical platelet) OR (glanzman s disease) OR (hereditary thrombasthenia) OR (platelet changes) OR (thromboasthenia) OR (thrombasthenia)
PubMed $_{Query}$	alcohols	*MeSH*	(("alcohols"[MH] OR ("alcohols"[TW])))
PubMed UmlsExt $_{Query}$	alcohols	*MeSH*	(("alcohols"[MH] OR ("alcohols"[TW] OR "alcohol, nos"[TW])))
PubMedInterExt $_{Query}$	alcohols	*MeSH*	(("alcohols"[MH] OR ("alcohols"[TW] OR "Ethanol"[TW] OR "Alcohol"[TW] OR "drinking"[TW] OR "alcohol consumption"[TW] OR "alcohol, nos"[TW])))
PubMed Manual $_{Query}$	allergens	*ATC*	"desensitization, immunologic"[MH] AND "allergens"[MH]"

Results

The InfoRoute Web Service

The web service can be used by third-party web services or web-based applications. For example, The HeTOP portal used the InfoRoute to access PubMed from any BMTO's terms. In the CISMeF catalog, a specific website link which can be refreshed after any CISMeF query is integrated. The website catagory corresponds to the "search engine category" from the twelve (12) catagories introduced above: PubMed, Intute, NCBI GQuery, etc.

InfoRoute Web Applications

Two web applications were devloped based on InfoRoute:

1. **The InfoRoute General Website Tool**

The application at http://inforoute.chu-rouen.fr/ir, visualises the input and the search results of the InfoRoute algorithm described in this paper. The input query allows the end-user to enter a term or an expression in French or English. All terminologies are assigned to the query by default. However, it is possible to specify the terminology(ies) input. Term or expression can be expressed using NI_{Query} or BI_{Query} described above. Search results can be obtained from all the twelve (12) category websites. Each category will be displayed seperately and websites in each category will be seperated according to their languages (French or English).

2. **The InfoRoute Clinical Trials Access**

The application at http://inforoute.chu-rouen.fr/irec, can be considered as a subtype of the InfoRoute general website tool described above. Tha aim of this application is to perform an effective access especially in French but also in English for

two clinical trial websites: http://clinicaltrials.gov (US clinical trials) and http://www.clinicaltrialsregister.eu (EU clinical trials). Additional options will be added to this applcaition compared to the general application based on AI_{Query} to genrate and $Advanced_{Query}$ described in the query translation section.

Discussion

The aim of this study was to present the InfoRoute algorithm, which uses multiple BMTO and multiple relationships to generate IR health queries from natural or Boolean queries in French or English. The InfoRoute algorithm permits to answer the two questions presented in the introduction:

- **Inadequate expression of information needs:** the use of multiple kinds of queries (natural, Boolean and advanced), it helps users to express their needs in multiple forms in two languages. Futhermore, the algorithm permits the user to express their queries according to multiple BMTO directly from specific biomedical terms or indirectly when applied the MTAIQ to extract biomedical terms.
- **Lack of the relevant information:** As described for PubMed and some clinical websites; the algorihm has been improved to retrieve the more accurate information for these specific websites. For example, in Thirion et al. [11] the optimzed query is significantly more precise than the current PubMed query (54.5% vs. 27%). In [12], the expansion of the PubMed query on UMLS synonyms increase recall and proportion of queries retrieving. However, the expansion of the user query according to multiple synonyms can decrease precision because the description of the same medical concept differ between BMTO. Besides UMLS synonyms, the $PubMedInterExt_{Query}$ which based on synonyms terms expanded using $CISMeF_{manual}$ and $CISMeF_{supervised}$ relationships will be evaluated in further work. Further sutdies will be perfomed to exploit hierachical inter relationships to expand results query.

The InfoRoute algorithm is the first algorithm developed which used French terms and/or French BMTO to generate and reformulate IR queries according to multiple health websites such as: PubMed or ClinicalTrials. The use of two languages is due to the bilingual BMTO included in HeTOP and thanks to multiple automatic and manual translations performed on multiple BMTO such as: MEDLINEPlus, MeSH SC, FMA, etc. Currenlty InfoRoute is integrated in the two main CISMeF tools: CISMeF search engine (http://doccismef.chu-rouen.fr/dc/) and HeTOP (http://www.hetop.eu/hetop/) to retrieve health infrormations from many sources. The next step is to integrate InfoRoute in real clinical information systems and develop a "French InfoButtons" as described in [14].

Conclusion

To conclude, the InfoRoute algorithm is usefuls tool to perform contextual information retrieval across multiple medical websites in both English and French.

Acknowledgments

This work was partially granted by: TerSan project (ANR TecSan program n°ANR-11-TECS-019-03). The ReLySe project (REfractory LYmphoma Sequencing) Transla 2013. The Authors are grateful to Richard Medeiros, for editing assistance.

References

[1] Podichetty V. Booher J., Whitfield M. Biscup R. A. Assessment of internet use and effects among healthcare professionals: a cross sectional survey. ost grad. Med. J. 2006; 2:2744-9.

[2] De Leo G, LeRouge C, Ceriani C, Niederman F. Websites most frequently used by physician for gathering medical information. AMIA Annual Symposium Proceedings, 2006; pp. 206.

[3] Freeman MK, Lauderdale SA, Kendrach MG, Woolley TW. Google Scholar versus PubMed in locating primary literature to answer drug-related questions. Ann Pharmacother 2009;43:478.

[4] Anders ME, Evans DP. Comparison of PubMed and Google Scholar literature searches. Respiratory care. 2010, 55:578-83.

[5] Braun L. Pro-active medical information retrieval. PhD Thesis, 2009.

[6] ter Hofstede AHM, Proper HA, van der Weide ThP. Query Formulation as an Information Retrieval Problem. The Computer Journal 1996; 39;pp. 255-274.

[7] Lindberg D, Humphreys B, McCray A. The Unified Medical Language System. Methods Inf Med. 1993;32(4):281–291.

[8] Grosjean J, Merabti T, Dahamna B, Kergouraly I, Thirion B, Soualmia L, Darmoni, SJ. Health Multi-Terminology Portal: a semantic added-value for partient safety. In: PSIP Workshop; 2011. p. 129–138.

[9] Merabti T, Soualmia LF, Grosjean J, Joubert M & Darmoni SJ. Aligning Biomedical Terminologies in French: Towards Semantic Interoperability in Medical Applications. In Book: Medical Informatics, pp. 41-68, InTech, 2012.

[10] Pereira S.: Multi-terminology indexing of concepts in health. [Indexation multi-terminologique de concepts en santé]. PhD Thesis, University of Rouen, Normandy,France.

[11] Darmoni SJ, Sakji S, Pereira S, Merabti T, Prieur E, Joubert M, Thirion B. Multiple terminologies in an health portal: automatic indexing and information retrieval. Artificial Intelligence in Medicine, Lecture Notes in Computer Science; 2009:255-259.

[12]Thirion B, Robu I, Darmoni, SJ. Optimization of the PubMed Automatic Term Mapping. Stud Health Technol Inform.2009;150: 238-242.

[13]Griffon N, Chebil W, Rollin L, Kerdelhue G, Thirion B, Gehanno JF, Darmoni SJ. Performance evaluation of Unified Medical Language System®'s synonyms expansion to query PubMed. BMC Medical Informatics and Decision Making. 2012;12(1):12.

[14]Merabti T, Abdoune H, Letord C, Sakji S, Joubert M, Darmoni SJ. Mapping the ATC classification to the UMLS Metathesaurus: some pragmatic applications. Patient Safety Informatics - Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety. Studies in Health Technology and Informatics, Volume 166, Pages 206-213, 2011.

[15]Darmoni SJ, Pereira S, Névéol A, Massari P, Dahamna B, Letrod C, Kerdeluhé G, Piot J, Derville A, Thirion B. French Infobutton: an academic and…business perspective. AMIA Annual Symposium Proceedings;pp. 920,2008.

Address for correspondence

Tayeb Merabti , CISMeF & TIBS, LITIS EA 4108, 1 rue de Germont
76031 Rouen Cedex, France, Tayeb.merabti@chu-rouen.fr.