We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900
Open access books available

185,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Chapter 3

Supporting E-Health Information Seekers: From Simple Strategies to Knowledge-Based Methods

Lina F. Soualmia, Badisse Dahamna and Stéfan J. Darmoni

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/50348

1. Introduction

Today, a web search is clearly one of the foremost methods for finding information. The growth of the Internet and the increasing availability of online resources have made the task of searching a crucial one. However, searching the web is not always as successful as users expect it to be and Internet users have to make a great effort to formulate a search query that returns the required results. Information retrieval concentrates on developing algorithms to locate and select documents from a corpus that are relevant to a given query. The development of online information retrieval tools, such as search engines or search robots many of which utilize hyperlink analysis [1], has been greatly beneficial to Internet users [2]. In the health domain, users are now experiencing huge difficulties in finding precisely what they are looking for among the numerous documents available online, and this in spite of existing tools. In medicine and health-related information accessible on the Internet, general search engines, such as Google, or general catalogues, such as Yahoo, cannot solve this problem efficiently [3]. This is because they usually offer a selection of documents that turn out to be either too large or ill-suited to the query. Free text word-based search engines typically return innumerable completely irrelevant hits, which require much manual weeding by the user, and also miss important information resources.

In this context, several health gateways [4] have been developed to support systematic resource discovery and help users find the health information they are looking for. These information seekers may be patients but also health professionals, such as physicians searching for clinical trials. Health gateways rely on thesauri and controlled vocabularies. Some of them are evaluated in [5]. Medical thesauri are a proven key technology for effective access to health information since they provide a controlled vocabulary for indexing documents and coding electronic health records. They therefore help to overcome some of the problems of free-text search by linking and grouping terms and concepts.

INTECH
open science | open minds

© 2012 Soualmia et al., licensee InTech. This is an open access chapter distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Nonetheless, medical vocabularies are difficult to handle by non-professionals. Problems also arise because there are practically as many different terminologies, controlled vocabularies, thesauri and classification systems as there are fields of application in health. We give in this chapter a panel of techniques that may be applied to help health information seekers. All the tests are performed on the CISMeF catalogue (Catalogue and Index of Medical Sites in French) [6] but are reproducible in other languages and other medical applications.

The remainder of the chapter is organized as follows: in section 2 we start by describing the CISMeF catalogue. The section 3 is devoted to simple search techniques such as approximate string matching and heuristics for queries composed by several words. Another method consists in meta-modeling health terminologies to improve information retrieval, the description of which is in the section 4. In the section 5 we describe the data-mining process to extract new knowledge and relations between terms to allow users to extend their searches.

2. The CISMeF catalogue

The CISMeF project was initiated in February 1995. As opposed to Yahoo, CISMeF is cataloguing the most important and quality-controlled sources of institutional health information in French. The CISMeF catalogue describes and indexes a large number of health-information resources of high quality (n=13,452 in October 2003; n=90,056 in May 2012). A resource can be a web site, web pages, documents, reports and teaching material: any support that may contain health information.

CISMeF takes into account the diversity of the end-users and allow them to find good quality resources. These resources are selected according to strict criteria by a team of librarians and are indexed according to a methodology which involves a four-fold process: resource collection, filtering, description and indexing. CISMeF is a quality-controlled gateway such as defined by Koch [4]. The following elements that characterize a typical quality-controlled health gateway are fulfilled in CISMeF: selection and collection development, collection management, intellectual creation of metadata, resource description (a metadata set), resource indexing (with controlled vocabulary system). To include only reliable resources, and to assess the quality of health information on the Internet, the main criteria (*e.g.* source, description, disclosure, last update) of CISMeF are from HONCode[1]. In the following sections we describe the set of metadata elements and the reference dictionary used in the catalogue.

2.1. CISMeF metadata

The notion of metadata was around before the Internet but its importance has grown with the increasing number of electronic publications and digital libraries. The World Wide Web Consortium (W3C) have proposed that metadata should be used to describe the data

[1] http://www.hon.ch/

contained on the web and to add semantic markup to web resources, thus describing their content and functionalities, from the vocabulary defined in terminologies and ontologies.

Metadata are data about data, and in the web context, these are data describing web resources. When properly implemented, metadata enhance information retrieval. The CISMeF uses several sets of metadata. Among them there is the Dublin Core (DC) [7] metadata set, which is a 15-element set intended to aid discovery of electronic resources. The resources indexed in CISMeF are described by eleven of the Dublin Core elements: *author, date, description, format, identifier, language, editor, type of resource, rights, subject* and *title.* DC is not a complete solution; it cannot be used to describe the quality or location of a resource. To fill these gaps, CISMeF uses its own elements to extend the DC standard. Eight elements are specific to CISMeF: *institution, city, province, country, target public, access type, sponsorships,* and *cost.* The user type is also taken into account. The CISMeF have defined two additional fields for resources intended for health professionals: indication of the *evidence-based medicine,* and the *method* used to determine it. For teaching resources, eleven elements of the IEEE 1484 LOM (Learning Object Metadata) "Educational" category are added.

2.2. CISMeF controlled vocabulary

Thesauri are a proven key technology for effective access to information as they provide a controlled vocabulary for indexing information. They therefore help to overcome some of the problems of free-text search by relating and grouping relevant terms in a specific domain. The main thesaurus used for medical information is the Medical Subject Headings (MeSH) [8] thesaurus used by the U.S. National Library of Medicine to index MEDLINE articles. The core of MeSH is a hierarchical structure that consists of sets of descriptors. At the top level we find general headings (*e.g.* diseases), and at deeper levels we find more specific headings (*e.g.* asthma). The 2012 version of the MeSH contains over 26,581 main headings (*e.g.* hepatitis, abdomen) and 83 subheadings (*e.g.* diagnosis, complications). Together with a main heading, a subheading allows to specify which particular aspect of the main heading is being addressed. For example, the pair [hepatitis/diagnosis] specifies the diagnosis aspect of hepatitis. For each main heading, MeSH defines a subset of allowable qualifiers so that only certain pairs can be used as indexing terms (*e.g.* aphasia/metabolism and hand/surgery are allowable, but hand/metabolism is not). The reference dictionary of CISMeF (the structure of which is detailed in Table 1) was created between 1995 and 2005 exclusively on the French version of the MeSH thesaurus maintained by the US National Library of Medicine, completed by numerous synonyms in French collected by the CISMeF team.

Several add-ons were performed around the MeSH thesaurus to index Web resources instead of scientific articles [9]: super-concepts (or Meta-terms) to optimize information retrieval and categorization, and resource types (organized hierarchically since 1997 *vs.* MeSH publication types' hierarchy since 2006). Indeed, MeSH main headings and subheadings are organized hierarchically but these hierarchies do not allow a complete view concerning a specialty. The main headings and subheadings in the CISMeF controlled vocabulary are brought together under metaterms (*e.g.* cardiology). Metaterms (n=73) concern

medical specialties and it is possible by browsing to know sets of MeSH main headings and subheadings qualifiers which are semantically related to the same specialty but dispersed in several trees. The MeSH thesaurus was originally used to index biomedical scientific articles for the MEDLINE database. In addition to the set of metaterms, the CISMeF team has modeled a hierarchy of resource types (n=127), to customize MeSH to the field of e-health resources. These resource types describe the nature of the resource (*e.g.* teaching material, clinical guidelines, patient forums), and are a generalization or extension of the MEDLINE publication types. Each resource in CISMeF is described with a set of MeSH main headings, subheadings and CISMeF resource types. Each main heading, [main heading/subheading] pair, and resource type is allotted a 'minor' or 'major' weight, according to the importance of the concept it refers to in the resource. Major terms are marked by a star (*).

	MeSH Terms	MeSH Synonyms	CISMeF synonyms	Total
1 word	9,679	9,391	3,359	22,429
2 words	9,833	28,051	8,258	46,142
3 words	4,204	19,551	6,569	30,324
4 words and +	2,503	16,992	4,924	24,419

Table 1. Composition of the reference dictionary based on the MeSH in French.

2.3. Searching through the catalogue

Many ways of navigation and information retrieval are possible in the catalogue [6]. The most used is the simple search (free text interface). It is based on subsumption relationships. If the query can be matched with an existing term of the terminology, thus the result is the union of the resources that are indexed by the term, and the resources that are indexed by the terms it subsumes, directly or indirectly, in all the hierarchies it belongs to. If the query cannot be matched, the search is done over the other fields of the metadata and in a worse case a full-text search is carried out. Contrary to MEDLINE, the resource types and the meta-terms were voluntary made ambiguous to maximize the recall (e.g. in the query guidelines in virology, virology will be recognized as a meta-term (instead of a term) and guidelines will be recognized as both the term and the resource type because we assume most of end users confuse content and container). In the following section we propose some simple enhancements for health information seekers' queries matching.

3. Spell-checking queries

A simple spelling corrector, such as Google's "*Did you mean:*" or Yahoo's "*Also try:*" feature may be a valuable tool for non-professional users who may approach the medical domain in a more general way [10]. Such features can improve the performance of these tools and provide the user with the necessary help. In fact, the problem of spelling errors represents a major challenge for an information retrieval system. If the queries (composed by one or multiple words) generated by information seekers remain undetected, this can result in a lack of outcome in terms of search and retrieval. A spelling corrector may be classified in

two categories. The first relies on a dictionary of well-spelled terms and selects the top candidate based on a string edit distance calculus. An approximate string matching algorithm, or a function, is required to detect errors in users' queries. It then recommends a list of terms, from the reference dictionary, that are similar to each query word. The second category of spelling correctors uses lexical disambiguation tools in order to refine the ranking of the candidate terms that might be a correction of the misspelled query.

3.1. Related work

Several studies have been published on this subject. We cite the work of Grannis [11] which describes a method for calculating similarity in order to improve medical record linkage. This method uses different algorithms such as Jaro-Winkler, Levenshtein [12] and the longest common subsequence (LCS). In [13] the authors suggest improving the algorithm for computing Levenshtein similarity by using the frequency and length of strings. In [14] a phonetic transcription corrects users' queries when they are misspelled but have similar pronunciation (*e.g.* Alzaymer *vs.* Alzheimer). In [15] the authors propose a simple and flexible spell-checker using efficient associative matching in a neural system and also compare their method with other commonly used spell-checkers. In fact, the problem of automatic spell checking is not new. Indeed, research in this area started in the 1960's [16] and many different techniques for spell-checking have been proposed since then. Some of those techniques exploit general spelling error tendencies and others exploit phonetic transcription of the misspelled term to find the correct term. The process of spell-checking can generally be divided into three steps:

i. error detection: the validity of a term in a language is verified and invalid terms are identified as spelling errors;
ii. error correction: valid candidate terms from the dictionary are selected as corrections for the misspelled term;
iii. ranking: the selected corrections are sorted in decreasing order of their likelihood of being the intended term.

Many studies have been performed to analyze the types and the tendencies of spelling errors for the English language. According to [17] spelling errors are generally divided into two types, (i) typographic errors and (ii) cognitive errors. Typographic errors occur when the correct spelling is known but the word is mistyped by mistake. These errors are mostly related to keyboard errors and therefore do not follow any linguistic criteria (58% of these errors involve adjacent keys [18] and occur because the wrong key is pressed, or two keys are pressed, or keys are pressed in the wrong order …*etc.*). Cognitive errors, or orthographic errors, occur when the correct spelling of a term is not known. The pronunciation of the misspelled term is similar to the pronunciation of the intended correct term. In English, the role of the sound similarity of characters is a factor that often affects error tendencies [18]. However, phonetic errors are harder to correct because they deform the word more than a single insertion, deletion or substitution. Damereau [16] indicated that 80% of all spelling errors fall into one of the following four single edit operation categories : (i) transposition of two adjacent letters (*as**h**tma* vs. *asthma*) (ii)

insertion of one letter (*asthm**m**a* vs. *asthma*) (iii) deletion of one letter (*as**tm**a* vs. *asthma*) and (iv) replacement of one letter by another (*asth**l**a* vs. *asthma*). Each of these wrong operations costs 1 *i.e.* the distance between the misspelled and the correct word [[17].

The third step in spell-checking is the ranking of the selected corrections. Main spell-checking techniques do not provide any explicit mechanism. However, statistical techniques [19] provide ranking of the corrections based on probability scores [20] with good results [21]. HONselect [22] is a multilingual and intelligent search tool integrating heterogeneous web resources in health. In the medical domain, spell-checking is performed on the basis of a medical thesaurus by offering information seekers several medical terms, ranging from one to four differences related to the original query. Exploiting the frequency of a given term in the medical domain can also significantly improve spelling correction [23]: edit distance technique is used for correction along with term frequencies for ranking. In [24] the authors use normalization techniques, aggressive reformatting and abbreviation expansion for unrecognized words as well as spelling correction to find the closest drug names within RxNorm for drug name variants that can be found in local drug formularies. It returns only drug name suggestions. To match queries with the MeSH thesaurus, Wilbur et al. [25] proposed a technique on the noisy channel model and statistics from the PubMed logs.

3.2. Proposed method

Research has focused on several different areas, from pattern matching algorithms and dictionary searching techniques to optical character recognition of spelling corrections in different domains. However, the literature is quite sparse in the medical domain, which is a distinct problem, because of the complexity of medical vocabularies. In this section, a simple method is proposed: it combines two approximate string comparators, the well-known Levenshtein [6] edit distance and the Stoilos function similarity defined in [26] for ontologies. We apply and evaluate these two distances, alone and combined, on a set of sample queries in French submitted to the health gateway CISMeF. A set of 127,750 queries were extracted from the query log server (3 months logs). Only the most frequent queries were selected. In fact some queries are more frequent than others. For example, the query *"swine flu"* is more present in the query log than *"chlorophyll"*. We eliminated the doubles (68,712 queries remained). From these 68,712 queries, we selected 25,000 queries to extract those with no answers (7,562). A set of 6,297 frequent queries was constituted from the original set of 7,562 by eliminating those that were submitted only once. In this set, the queries were composed from 1 to 4 and more words as detailed in the Table 2.

Composition	Number
1 word	1,061
2 words	1,636
3 words	1,443
4 (and more) words	2,157
Total	6,297

Table 2. Structure of the queries (with no answer) obtained from the logs.

3.2.1. Similarity functions

Similarity functions between two text strings S_1 and S_2 give a similarity or dissimilarity score between S_1 and S_2 for approximate matching or comparison. For example, the strings "*Asthma*" and "*Asthmatic*" can be considered similar to a certain degree. Modern spell-checking tools are based on the simple Levenshtein edit distance [12] which is the most widely known. This function operates between two input strings and returns a score equivalent to the number of substitutions and deletions needed in order to transform one input string into another. It is defined as the minimum number of elementary operations that is required to pass from a string S_1 to a string S_2. There are three possible transactions: replacing a character with another, deleting a character and adding a character. This measure takes its values in the interval [0, ∞ [. The Normalized Levenshtein [27] (*LevNorm*) in the range [0,1] is obtained by dividing the distance of Levenshtein $Lev(S_1, S_2)$ by the size of the longest string and it is defined by the following equation ():

$$\text{LevNorm}(S_1, S_2) = \frac{\text{Lev}(S_1, S_2)}{\text{Max}(|S_1|, |S_2|)} \quad (1)$$

For example, LevNorm(eutanasia, euthanasia)=0.1, as Lev(eutanasia, eut**h**anasia)=1 (adds 1 character h); |eutanasia|=9 and |euthanasia|=10.

We complete the calculation of the Levenshtein distance by the similarity function Stoilos proposed in [26]. It has been specifically developed for strings that are labels of concepts in ontologies. It is based on the idea that the similarity between two entities is related to their commonalities as well as their differences. Thus, the similarity should be a function of both these features. It is defined by the equation (2) where $Comm(S_1,S_2)$ stands for the commonality between the strings S_1 and S_2, $Diff(S_1,S_2)$ for the difference between S_1 and S_2, and $Winkler(S_1,S_2)$ for the improvement of the result using the method introduced by Winkler in [28]:

$$\text{Sim}(S_1, S_2) = Comm(S_1, S_2) - Diff(S_1, S_2) + winkler(S_1, S_2) \quad (2)$$

The function of commonality is determined by the substring function. The biggest common substring between two strings (*MaxComSubString*) is computed. This process is further extended by removing the common substring and by searching again for the next biggest substring until none can be identified. The function of commonality is given by the equation (3):

$$\text{Comm}(S_1, S_2) = \frac{2 \times \sum_i |\text{MaxComSubString}_i|}{|S_1| + |S_2|} \quad (3)$$

For example, for S_1=**Trigonocep**ah**lie** and S_2=**Trigonocep**ha**lie** we have: |MaxComSubString$_1$| = |Trigonocep|=10, |MaxComSubString$_2$| =|lie|=3 and *Comm*(Trigonocepahlie,Trigonocephalie) = 0.866.

The difference function *Diff(S_1,S_2)* is based on the length of the unmatched strings resulting from the initial matching step. The function of difference is defined in equation (4) where $p \in [0, \infty[$, $|u_{S1}|$ and $|u_{S2}|$ represent the length of the unmatched substring from the strings S_1 and S_2 scaled respectively by their length :

$$\text{Diff}(S_1, S_2) = \frac{|u_{S1}| \times |u_{S2}|}{p + (1-p) \times (|u_{S1}| + |u_{S2}| - |u_{S1}| \times |u_{S2}|)} \quad (4)$$

For example for S_1=Trigonocepahlie and S2=Trigonocephalie and p=0.6 we have: $|u_{S1}|$ = 2/15; $|u_{S2}|$=2/15; *Diff(S_1,S_2)* =0.0254.

The Winkler parameter *Winkler(S_1,S_2)* is a factor that improves the results. It is defined by the equation (5) where *L* is the length of common prefix between the strings S_1 and S_2 at the start of the string up to a maximum of 4 characters and *P* is a constant scaling factor for how much the score is adjusted upwards for having common prefixes. The standard value for this constant in Winkler's work is *P=0.1* :

$$\text{Winkler}(S_1, S_2) = L \times P \times (1 - Comm(S_1, S_2)) \quad (5)$$

For example, for between S_1=hyperaldoterisme and S_2=hyperaldosteronisme, we have $|S_1|$=16, $|S_2|$=19; the common substrings between S_1 and S_2 are hyperaldo, ter, and isme. Comm(S_1,S_2)=0.914; Diff(S_1,S_2)=0; Winkler(S_1,S_2)=0.034 and Sim(**hyperaldoterisme**,**hyperaldoster**on**isme**)=0.948.

3.2.2. Processing users' queries

As detailed in [18], spelling errors can be classified as typographic and phonetic. Cognitive errors are caused by a writer's lack of knowledge and phonetic ones are due to similar pronunciation of a misspelled and corrected word. We pre-process the queries by a phonetic transcription with the algorithm described in [14]. To process multi-word queries, we used the following basic natural language processing steps and the well-known Bag-of-Words (BoW) algorithm before applying similarity functions:

1. *Query segmentation*: the query was segmented in words thanks to a list of segmentation characters and *string tokenizers*. This list is composed of all the non-alphanumerical characters (*e.g.*: * $,!§;|@).
2. *Character normalizations*: we applied two types of character normalization at this stage. MeSH terms are in the form of non-accented uppercase characters. Nevertheless, the terms used in the CISMeF terminology are in mixed-case and accented. (1) *Lowercase conversion*: all the uppercased characters were replaced by their lowercase version; "*A*" was replaced by "*a*". This step was necessary because the controlled vocabulary is in lowercase. (2) *Deaccenting*: all accented characters (*"éèêë"*) were replaced by non-accented (*"e"*) ones. Words in the French MeSH were not accented, and words in queries were either accented or not, or wrongly accented (*hèpatite"* instead *"hépatite"*).

3. *Stop words*: we eliminated all stop words (such as *the, and, when*) in the query. Our stop word list was composed 1,422 elements in French (*vs.* 135 in PubMed).
4. *Exact match expression*: we use regular expressions to match the exact expression of each word of the query with the terminology. This step allowed us to take into account the complex terms (composed of more than one word) of the reference dictionary and also to avoid some inherent noise generated by the truncations. The query *'**accident**'* is matched with the term *'circulation **accident**'* but not with the terms *'**accident**s'* and *'chute **accident**elle'*. The query *'**sida**'* is matched with the terms *'lymphome lié **sida**'* and *'**sida** atteinte neurologique'* but not with the terms *'gluco**sida**ses'*, *'agra**sida**e'* and *'bêta galacto**sida**se'* which are not relevant.
5. *Phonemisation:* It converts a word into its French phonemic transcription: *e.g.* the query *alzaymer* is replaced by the reserved term *alzheimer*.
6. *Bag of words:* The algorithm searched the greatest set of words in the query corresponding to a reserved term. The query was segmented. The stop words were eliminated. The other words were transformed with the *Phonemisation* function and sorted alphabetically. The different reserved term bags were formed iteratively until there were no possible combinations. The query *'therapy of the breast cancer'* gave two reserved words: '*therapeutics'* and '*breast cancer'* (*therapy* being a synonym of the reserved term *therapeutics).*

3.2.3. Evaluations

To evaluate our method of correcting misspellings, we used the standard measures of evaluation of information retrieval systems, by calculating precision, recall and the F-Measure. We performed a manual evaluation to determine these measures. Precision (6) measured the proportion of queries that were properly corrected among those corrected.

$$Precision = \frac{\left|\{Queries\ correctly\ corrected\}\right|}{\left|\{Queries\ corrected\}\right|} \tag{6}$$

Recall (7) measured the proportion of queries that were properly corrected among those requiring correction.

$$Recall = \frac{\left|\{Queries\ correctly\ corrected\}\right|}{\left|\{Queries\ to\ be\ corrected\}\right|} \tag{7}$$

The F-Measure combined the precision and recall by the following equation (8) :

$$F - Measure = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \tag{8}$$

We also calculated confidence intervals at ρ=5% to avoid evaluating the whole set of queries, but some sets that are manually manageable. For a proportion x and a set of size n_x the confidence interval is:

$$CI_x = \left[x - 1.96 \times \sqrt{\frac{x \times (1-x)}{n_x}}; x + 1.96 \times \sqrt{\frac{x \times (1-x)}{n_x}} \right] \quad (9)$$

3.2.4. Results

The Levenshtein and Stoilos functions require a choice of thresholds to obtain a manageable number of correction suggestions for the user. We tested, in a previous work, different thresholds [29] for the normalized Levenshtein distance, the similarity function of Stoilos and for the combination of both on a set of 163 queries. The best results were obtained with Levenshtein>0.2 and Stoilos>0.7. To determine the impact of the size of the query we measured the number of suggestions of corrected queries (on the set of 6,297 frequent queries) in the Table 3. For a user, the maximum number of manageable suggestions for one query was 6.

	Nb characters	Nb suggestions by query
1 word query	Min = 3; Avg = 10.49 ; Max = 25	Avg = 0.39 ; Max = 5
2 words query	Min = 5; Avg = 18.36; Max = 41	Avg = 0.22 ; Max = 6
3 words query	Min = 10; Avg = 24.39; Max = 54	Avg = 0.13; Max = 1
4 words and +query	Min = 11; Avg = 37.30; Max = 113	Avg = 0.06; Max = 1

Table 3. Number of suggestions according to the size of the queries.

Manual evaluations were performed on sets of ~1/3 of each type of queries. Evaluations of the quality of queries suggestions (Precision, Recall and F-Measure) were performed manually on several sets, according to the size of the query, but also according to the following methods : Bag-of-Words, Levenshtein distance alongside the Stoilos similarity function, but also the Bag-of-Words processed before and after the combination of the Levenshtein distance along with the Stoilos similarity function. Levenshtein and Stoilos remained constant at <0.2 and >0.7 respectively. The resulting curves are in Figures 1, 2 and 3. By combining the Bag-of-Words algorithm along with the Levenshtein distance and the similarity function of Stoilos, a total of 1,418 (22.52 %) queries matched medical terms or combinations of medical terms. The remaining queries with no suggestions (when terms and also the possible combination of terms) not belong to the dictionary. For 1-word queries, it remained 711 (67%), for 2-words queries it remained 1197 queries (73.16%); for 3-words queries it remained 1126 (78.08%) and for 4 words queries it remained 1,846 queries (85.58%). For example, the query "*nutrithérapie*" (nutritherapy) contains no error but cannot be matched with any medical term in the reference dictionary. Evaluations shown that best results were obtained by performing the Bag-of-Words algorithm before the combination of Levenshtein alongside Stoilos.

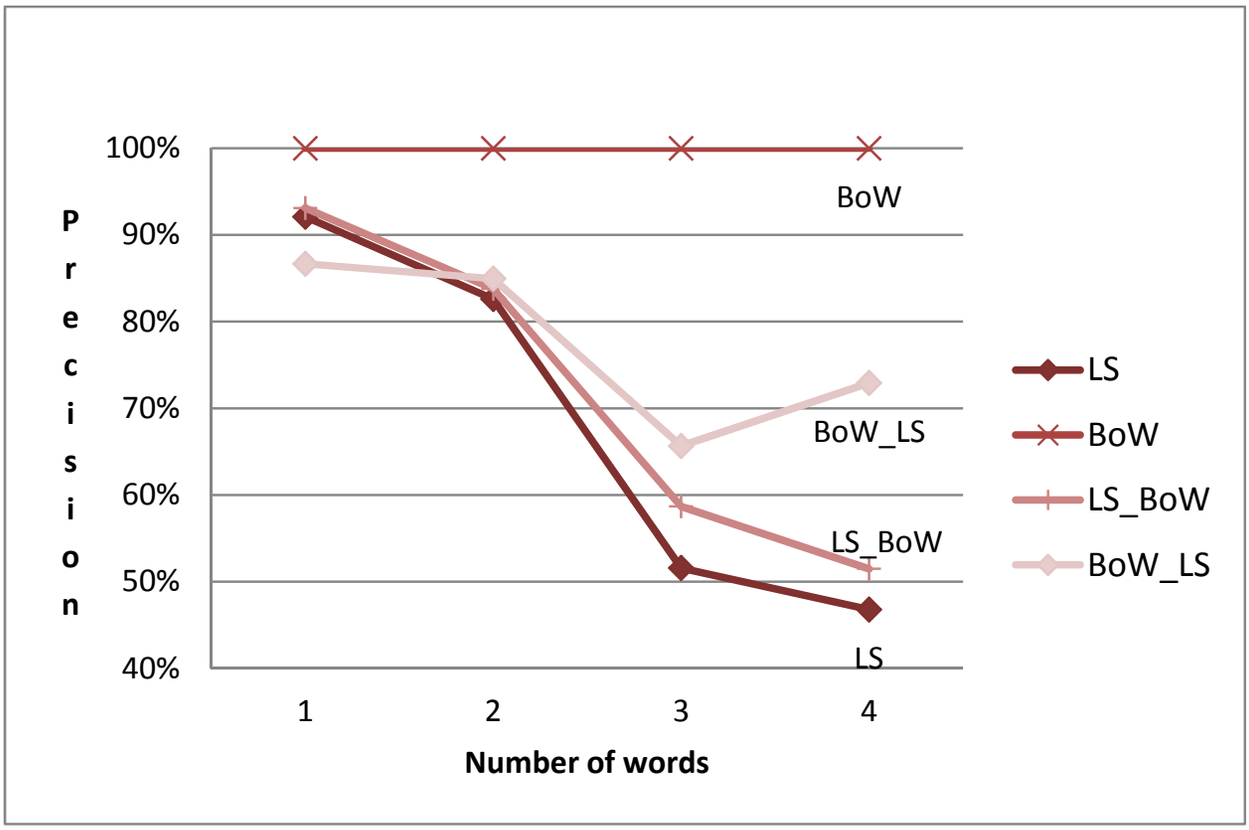


Figure 1. Precision curves according to the size of the query.

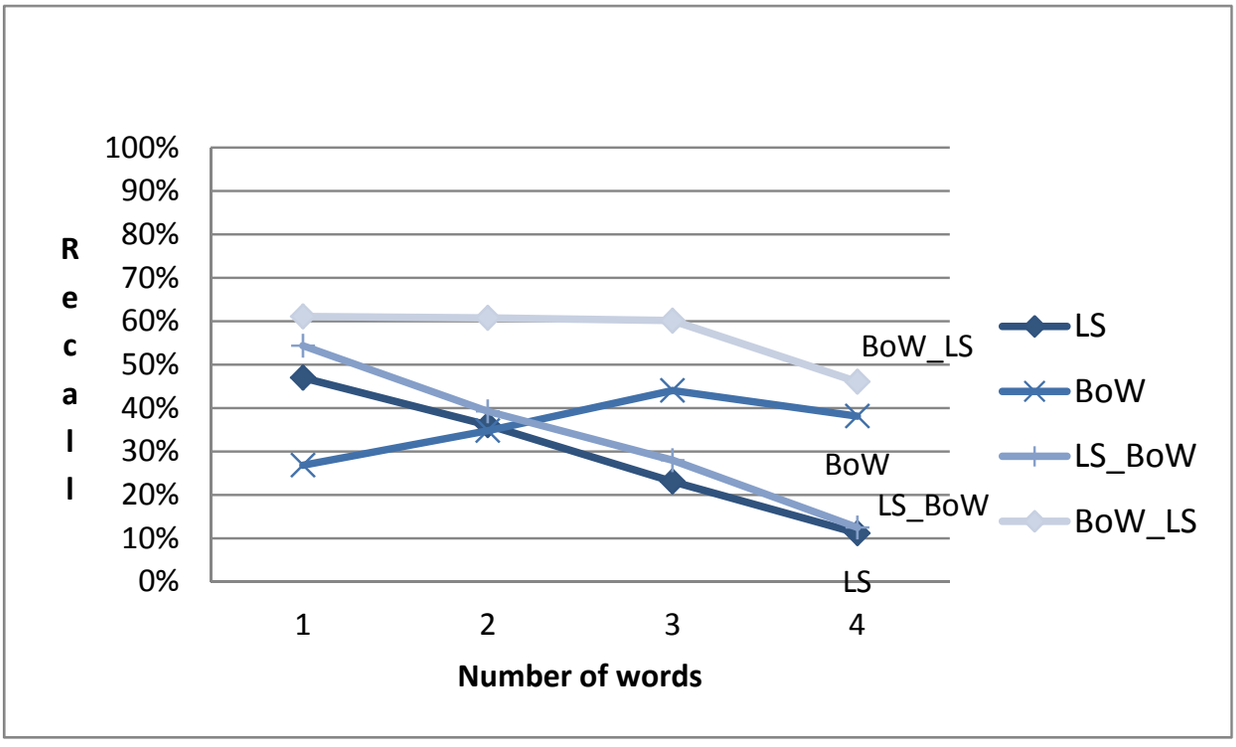


Figure 2. Recall curves according to the size of the query.

The different experiments we performed show that with 38% recall and 42% precision, *Phonemisation* cannot correct all errors : it can only be applied when the query and entry term of the vocabulary have similar pronunciation. However, when there is reversal of characters in the query, it is an error of another type: the sound is not the same and similarity distances such as Levenshtein and Stoilos can be exploited here. Similarly, when using certain characters instead of others ("*ammidale*" instead of "*amygdale*"), string similarity functions are not efficient. The best results (F-measure 64.18%) are obtained with multi-word queries by performing the Bag-of-Words algorithm first and then the spelling-correction based on similarity measures. Due to the relatively small number of correction suggestions (min 1 and max 6), which are manually manageable by a health information seeker, we have chosen to return an alphabetically sorted list rather than ranking them.

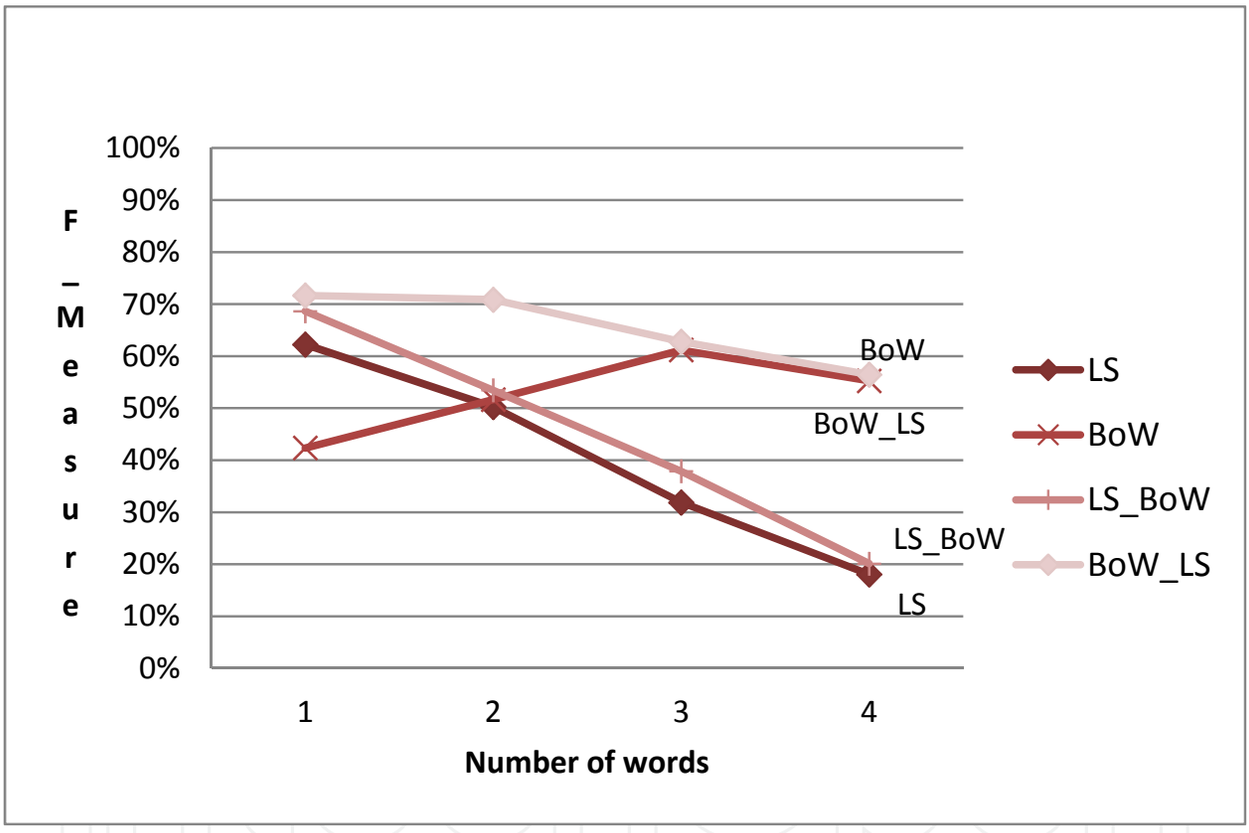


Figure 3. F-Measure curves according to the size of the query.

3.3. Simple heuristics

The complex terms matching is more requiring than simple terms matching. The CISMeF team editorial policy concerning the queries' rewriting consists in maximizing as much as possible the Doc'CISMeF recall. This approach is mainly due to the size of the CISMeF's corpus (n=90,056 *vs.* several million in the MEDLINE database). When all the terms of the query couldn't be recognized as reserved terms or couldn't be corrected by our spell-checker, we have implemented 5 main heuristics:

Step 1. *The reserved terms*: The process consists in recognizing the user query expression. If it matches a reserved term of the terminology, the process stops, and the answer of the query is the union of the resources that are indexed by the term, and the resources that are indexed by the terms it subsumes, directly or indirectly, in all the hierarchies it belongs to. If it doesn't match a reserved term, the query is segmented into seek if it contains one or more reserved terms. The query *'enfant asthme'* is replaced by the Boolean query (*enfant.mr AND asthme.mr*), where *enfant* and *asthme* are reserved terms (*mr*). The reserved terms are matched thanks to the bag of words algorithm independently of the words query order.

Step 2. *The documents' title*: The search is performed over the other fields of the metadata. The title of the documents is considered in priority. The stop words are eliminated and the search is realized over the union of the words of the query with a truncation (*) at the right in the field title (*ti*), as the following: $word_1$*.*ti AND* $word_2$*.*ti* for a 2-words query.

Step 3. *Mixing the reserved terms and the titles*: The system seeks if some words are reserved terms or not. A new Boolean query is generated with the fields reserved term (*mr*), if the word is a reserved term, and title (*ti*) if not. The query *'allergie infantil*e' is replaced by the Boolean query (*allergie.mr AND infantile.ti*).

Step 4. *Mixing the reserved terms, all fields and adjacency in the titles* : The search is processed over all the fields (*tc*) of the documents' metadata for the words that couldn't be recognized as reserved terms UNION the initial query processed over all the fields with adjacency (*at*) at *n* words with *n=5*(nb words of the query-1)*. The query *'les problems respiratoires des enfants*' is replaced by the Boolean query *[(enfant.mr AND problemes.tc AND respiratoires.tc) OR (problemes respiratoires enfant.at)]*. In this query, the word *enfant* is recognized as a reserved term because it has the same sonority as the reserved term *enfants*. The words *problèmes* and *respiratoires* are searched over all the fields and the initial query *problèmes respiratoires enfants* is searched over all the fields with adjacency of 10 which means that these 3 words shouldn't be distant at more than 10 words.

Step 5. *Mixing the reserved terms, all fields and adjacency in the plain texts* : A plain text search over the documents with adjacency (ap) of n words with n=10*(nb words of the query-1) is realized. The query 'bronchite asthmatiforme' is replaced by the Boolean query (bronchite asthmatiforme.ap) where the words bronchite and asthmatiforme shouldn't be distant at more than 10 words in the plain texts of the documents.

An intuitive scale of interpretation (from Step 1 to Step 5) is available to inform the users about their queries operations and rewritings. By using these simple heuristics, 65% of the queries returned documents (27% by the step 1; 7% by the step 2; 4% by the step 3; 10% by the step 4 and 17% by the step 5).

We describe in the next section how to maximize information retrieval by meta-modeling. The relevance on using multiple medical terminologies to improve information retrieval versus only the MeSH thesaurus is also evaluated.

4. Meta-modeling

To maximize information retrieval through the catalogue, one another enhancement is to gather all the MeSH terms that are related to a given specialty, since they can be dispersed among the 16 MeSH branches. On the other hand, the use of multiple terminologies is recommended [29] to increase the number of the lexical and graphical forms of a biomedical term recognized by a search engine. Since 2007, the CISMeF resources are indexed using the vocabulary of 23 other terminologies and classifications, most of them being bilingual (English and French). To supply health information seekers with the terminologies available in French, these terminologies are accessible through the Health Multiple Terminologies and Ontologies Portal (HeTOP) [31].

4.1. MeSH meta-terms for information retrieval

The MeSH thesaurus is partitioned at its upper level into 16 branches (*e.g.* Anatomy, Diseases). The core of MeSH thesaurus is a hierarchical structure that consists of sets of descriptors. However, these hierarchies do not allow a complete view concerning a specialty. The main headings and subheadings in the CISMeF controlled vocabulary are gathered under meta-terms (*e.g.* cardiology) (Figure 4). Meta-terms (n=73) concern medical specialties and it is possible by browsing to know sets of MeSH main headings and subheadings which are semantically related to the same specialty but dispersed in several trees. Meta-terms have been created to optimize information retrieval in CISMeF and to overcome the relatively restrictive nature of MeSH headings. For example a search on "guidelines" or "virology", where cardiology and virology are descriptors, yield few answers. Introducing cardiology and virology as meta-terms is an efficient strategy to obtain more results because instead of exploding one single MeSH tree, the use of meta-terms results in an automatic expansion of the queries by exploding other related MeSH trees besides the current tree, using the well-known automatic query expansion process. In other words, a query using a meta-term corresponds to the union of all the queries for all the terms semantically linked to it. A comparison of the results of MeSH term-based queries and SC-based queries showed an increased recall with no decrease in precision [33].

4.2. Multiple-terminologies meta-terms

The use of multiple terminologies is recommended [29] to increase the number of the lexical and graphical forms of a biomedical term recognized by a search engine. For this reason, CISMeF evolved recently from a single terminology approach using the MeSH main headings and subheadings to a multiple terminologies paradigm using, in addition to the MeSH thesaurus, vocabularies and classifications that deal with various aspects of health. Among them, the Systematized NOmenclature of MEDicine (SNOMED 3.5), the French CCAM for procedures [34], Orphanet for rare diseases[2] and some classifications from the World Health Organization : the 10th revision of the International Classification of Diseases[3]

[2] www.orpha.net

[3] http://www.who.int/classifications/icd/en/

(ICD10), Anatomical Therapeutic Chemical (ATC) Classification for drugs , ICF for handicap, ICPS for patient safety, MedDRA[4] for adverse effects. These terminologies were fully integrated into the CISMeF back-office. They can be used for indexing resources (allowing a more precise indexing) and thus for querying the catalogue. However, the addition of multiple terminologies to CISMeF did not induce modifications in the tasks performed for using, maintaining and updating the catalogue. The richest source of biomedical terminologies, thesauri, classifications is constituted by the Unified Medical Language System (UMLS) Metathesaurus initiated in by the U.S. NLM with the purpose to integrate information from a variety of sources. Nonetheless, the Metathesaurus does not allow interoperability between terminologies since it integrates the various terminologies as they stand without making any connection between the terms in the terminologies other than by linking equivalent terms to a single identifier in the Metathesaurus. The approach in CISMeF has the advantage of combining respect for the original structure of each of the terminologies with a re-grouping of the meta-data inherent in each terminology.

New terminologies have been linked to meta-terms manually by experts in CISMeF: one physician for ICD10, which is partitioned into 22 chapters, and the CCAM; one pharmacist-librarian for ATC, and one medical resident for the terms of the Foundational Model of Anatomy. For instance, the meta-term "*cardiology*" was initially linked to MeSH main headings such as "*cardiology*", "*stents*", and their descendants. With the integration of new terminologies, additional links completed the definition of the meta-term "*cardiology*": links to "*cardiovascular system*", "*Antithrombotic agents*" and others from ATC, links to "*Cardio-myopathy*", "*Heart*" and their descendants from ICD10 and so on.

4.2.1. Test queries

Our aim is to compare the precision and recall of multiple terminologies meta-terms (mt-mt) to MeSH meta-terms (M-mt) in CISMeF. Since mt-mt are based on M-mt plus semantic links to some terms in other terminologies, the query results for M-mt are all included in the query results for mt-mt, which became the gold standard for recall. We have then to evaluate the precision of the query retrieving resources indexed by a term linked to M-mt (MeSH meta-term query), on the one hand, and by a term linked to mt-mt and not to M-mt (Δ query) on the other hand. For this purpose, we build Boolean queries using the meta-terms themselves. For example, for the "*surgery*" meta-term, the MeSH meta-term (M-mt) query is "*surgery[M-mt]*". The Δ query is: "*surgery[mt-mt] NOT surgery[M-mt]*". Retrieved resources returned were assessed for relevance. We detail in the next section the criteria we have used for evaluation.

4.2.2. Evaluations

The resources returned by the CISMeF's search tool using automatic query expansion were assessed for relevance according to a three modality scale used in other standard Information Retrieval test sets: irrelevant (0), partly relevant (1) or fully relevant (2). A physician manually assigned relevance scores (0;1;2) to the top 20 resources returned for

[4] http://www.meddramsso.com

each meta-term query. The results of the evaluation are given in the Table 4. We chose to assign relevance scores to the top twenty resources returned because 95% of the end-users do not go beyond this limit when using a general search engine [35]. For the purpose of assessing meta-terms for Information Retrieval, we have developed a test collection comprising relevance judgments for the top 20 resources returned for a selection of 20 eta-terms queries. Table 4 shows that the queries yielded 118,772 resources, of which 708 were assessed for relevance (0.6%). Weighted precisions for MeSH meta-terms queries and for Δ queries were computed given the level of relevance considered and compared using χ^2 test. Indexing methods and meta-terms were compared too. Relative recall for MeSH meta-terms queries were computed given the level of relevance considered.

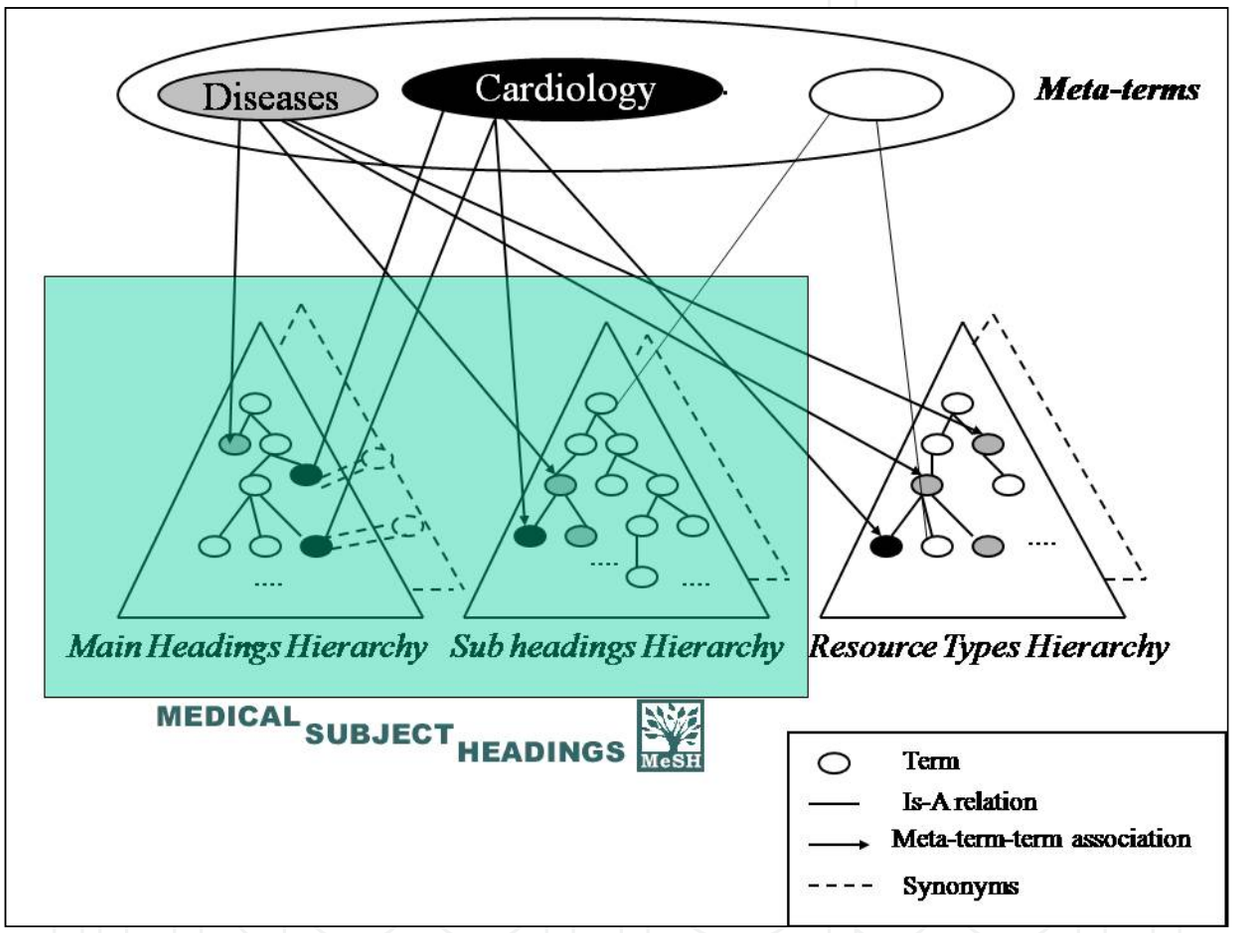


Figure 4. Gathering MeSH *main headings* and *subheadings* under *meta-terms*. *Resource types* are modelled to describe the nature of a resource because of the heterogeneity of resources.

The mean weighted precision of Δ queries was 0.33 and 0.76 for, respectively, full and partial relevance. The mean precision of MeSH meta-terms queries was 0.66 and 0.80 for, respectively, full and partial relevance. The difference between MeSH meta-terms and multiple terminologies meta-terms was significant for full relevance (0.66 *vs* 0.61; $p<10^{-4}$, χ^2) but not for partial relevance (both 0.80; p=0.3, χ^2). The mean recall of MeSH meta-terms queries was 0.92 and 0.86 for, respectively, full and partial relevance. Table 5 shows that, whatever the relevance considered was, results varied significantly according to the

indexing method: manual (precision of 0.50 and 0.81 for, respectively, full and partial relevance) perform better than automatic (precision of 0.38 and 0.48 for, respectively, full and partial relevance), and to the studied meta-term.

Meta-Term	Query Type	Nb of documents	Relevance on 20 doc		
			Not	Partially	Totally
Diagnosis	MeSH	13,132	0	2	15
	Delta	350	14	1	5
Toxicology	MeSH	11,980	0	0	20
	Delta	482	16	1	3
Neurology	MeSH	9,325	8	4	8
	Delta	2,168	11	5	4
Infectious Diseases	MeSH	6,557	0	0	20
	Delta	2,573	3	16	1
Paediatrics	MeSH	7,560	4	4	12
	Delta	251	2	4	13
Cardiology	MeSH	5,288	1	0	18
	Delta	2,388	4	10	6
Oncology	MeSH	5,626	0	1	18
	Delta	1,063	2	14	4
Surgery	MeSH	5,504	17	0	3
	Delta	320	5	0	15
Rheumatology	MeSH	4,408	3	8	9
	Delta	856	11	5	4
Gastroenterology	MeSH	4,069	0	0	20
	Delta	1,106	8	11	1
Allergies and Immunology	MeSH	4,598	1	17	2
	Delta	573	2	17	1
Metabolism	MeSH	3,797	14	2	4
	Delta	849	0	2	18
Dermatology	MeSH	3,196	7	0	13
	Delta	1,427	0	4	16
Nutrition	MeSH	3,455	0	1	19
	Delta	1,027	0	9	11
Pneumology	MeSH	3,466	0	7	12
	Delta	584	0	14	6
Gynaecology	MeSH	3,186	6	1	12
	Delta	850	0	1	19
Obstetrics	MeSH	3,063	5	1	12
	Delta	316	20	0	0
Virology	MeSH	3,122	1	11	6
	Delta	257	0	20	0
Total	**MeSH**	**101,332**	**67**	**59**	**223**
	Delta	**17,440**	**98**	**134**	**127**

Figure 5. Relevance of resources retrieved by 18 meta-terms queries on top 20 documents.

Variable	Full Relevance	Partial Relevance
Specific Query (M-mt vs mt-mt)	p < 10-4	p = 0.3
Indexing Method	p = 0.004	$p < 10^{-4}$
Meta-Term	p < 10-4	$p < 10^{-4}$

Figure 6. Determinants of relevance; χ^2 test.

To complete the information retrieval process and to allow interactive query expansion with the health information seeker, we propose in the next section to use "new" knowledge represented as association rules extracted by data-mining process.

5. Knowledge extraction

The knowledge-approach is based upon a data-mining process, called association rules, which can infer "new" relations between medical concepts. A data-mining system may generate several thousands and even several millions frequent association rules, and only some of these will be interesting. In this section we will show how only the most relevant association rules are mined using Formal Concept Analysis and Galois closure. We consider a relevant association rule as being non-redundant with a minimal antecedent and a maximal consequent, which is particularly useful for query expansion.

5.1. Association rules

The discovery of association rules is a widely used technique in data-mining. The general problem was described in [36], in which relations were discovered among pieces of data (called items). An association rule is interesting if it is easily understood by the users, valid for new data, useful, or confirms a hypothesis. The task of association rule mining can be applied to various types of data: any data set containing multiple items.

5.1.1. Definitions

Let I be a set of items, called itemset, and D a database of transactions where each transaction T ($T \in D$) is an itemset. An association rule is an implication rule expressed in the form of: $I_1 \rightarrow I_2$ where I_1 and I_2 are two itemsets $I_1, I_2 \subseteq I$ so that $I_1 \cap I2 = \varnothing$. The rule expresses that whenever a transaction T contains I_1 then T probably also contains $I2$. In other words, the implication rule means that the apparition of the itemset I_1 in a transaction T, implies the apparition of the itemset I_2 in the same transaction. However, the reciprocal implication does not have to happen necessarily. I_1 is called antecedent and I_2 is called consequent.

5.1.2. Support

The support of an association rule represents its utility. This measure corresponds to the proportion of objects which contains at the same time the rule antecedent and consequent. It

is possible to calculate the support of an association rule from the support of an itemset. Supp(I_k) the support of the itemset I_k is defined as the probability of finding I_k in a transaction of T:

$$Supp(I_k) = \frac{\left|\{t \in T / I_k \subseteq t\}\right|}{\left|T\right|} \tag{10}$$

The support of the rule $I_1 \rightarrow I_2$ written as Supp($I_1 \rightarrow I_2$) is calculated as follows:

$$Supp(I_2 \rightarrow I_1) = Supp(I_1 \cup I_2) \tag{11}$$

5.1.3. Confidence

The confidence of an association rule represents its precision. This measure corresponds to the proportion of objects that contains the consequent rule among those containing the antecedent. The confidence of the rule $I_1 \rightarrow I_2$, written as Conf($I_1 \rightarrow I_2$) is calculated as follows:

$$Conf(I_1 \rightarrow I_2) = \frac{Supp(I_1 \cup I_2)}{Supp(I_1)} \tag{12}$$

Two types of rules are distinguished: exact association rules that have a confidence equal to 100%, *i.e.* verified in all the objects of the database and approximate association rules that confidence<100%.

5.2. Data-mining algorithms

Several methods are used to extract all of the association rules from a database. The simplest method consists of enumerating all the itemsets from which all the possible association rules could be generated. The total number of itemsets for a database that contains n Boolean attributes is 2^n. This naïve method is inapplicable to real-life databases. A more efficient method involves computing itemsets that have a support higher than a given threshold. They are called *frequent itemsets*. The association rules extraction time depends on the frequent itemsets extraction time. Several accesses to the database are necessary to compute the number of database objects in which each frequent itemset candidate is contained. The association rules algorithms by level consider in each iteration a set of itemsets of a particular size, *i.e.* a set of itemsets in a level of the itemsets lattice. The following properties are used by these algorithms to limit the number of the itemsets candidates: all of the super-sets of an infrequent itemset are infrequent, and all the subsets of a frequent itemset are frequents [37]. This method is founded on the two-stepped model that finds all of the rules that satisfy user-specified minimum support and confidence: (i) Generate all large itemsets that satisfy minimum support and (ii) From large itemests generate all association rules that satisfy minimum confidence. Apriori algorithm [37] realizes a number of database accesses equal to the size of the larger

frequent itemsets. Many researchers have tried to improve various aspects of Apriori, such as the number of passes and accesses to the data-bases or the time efficiency of those passes. We have chosen to adapt the A-Close algorithm [38] in which new bases for association rules are deduced from the closed frequent itemsets and their generators. These bases consist of non-redundant association rules of minimal antecedents and maximal consequents, *i.e.* the most relevant association rules and are defined by using the closure operator of the Galois connection of a finite binary relation. All frequent itemsets and their support, and therefore all association rules, are deduced efficiently from the frequent closed itemsets without accessing the database.

5.3. Extracting knowledge from e-Health documents

Our experiments are carried out on the CISMeF database. An extraction context is a triplet *C= (O, I, R)* where *O* is the set of objects, *I* is the set of all the items and *R* is a binary relation between *O* and *I*. Applying this model to our database, the objects are the indexed e-health documents. Each document has a unique identifier and a set of associated descriptors. These descriptors may be MeSH main headings and associations between MeSH main headings and MeSH subheadings. The relation *R* represents the indexing relation between an object and an item, *i.e.* a descriptor that belongs to *I*. We studied different extraction contexts by applying and adapting the A-Close algorithm such as the context of categorized documents, according to the user type and to meta-terms. There is an average of 6.5 descriptors by document in CISMeF with a minimum of 1 and a maximum of 300. This constraint on the number of descriptors *i.e.* the size of the set of items has been considered in the implementation phase of the A-Close algorithm. Indeed, A-Close works on databases with a maximum of 12 items. We have added another requirement to the implementation to avoid long time generation: maximal size of the closed itemsets is fixed to 300 items as it corresponds to the maximum number of descriptors for the documents. As an output, the association rules may be visualized in a file or automatically added to the database to be used in the information retrieval process, mainly by interactive query expansion.

5.3.1. Extracting knowledge from all the database

- *Case 1*: In the first case, let *I* be the set of main headings (MH), which, via *R*, are used to index a subset *O* of 11,373 documents. The 11,373 documents were selected at random. We have fixed the support threshold as minsup=20 and the confidence threshold as minconf=70%. A total of 11,819 rules were mined (2,438 exact with confidence=100%; 9,381 approximate with confidence≥70%). The number of rules is too high to be manually analyzed by our experts (physicians or medical librarians).
- *Case 2*: In the second case, let I be the set of main headings (MH) and subheadings (SH) associated with the set of documents O. $I=\{MH\} \cup \{SH\}$. We obtained 16,976 rules (5,241 exact; 11,738 approximate). The same conclusions are drowned from the case 1 : too numerous rules to be evaluated manually.

- *Case 3*: In the third case, *I* is the set of the associations of main headings and subheadings (MH/SH) related to the documents. *I={[MH/SH]}*. Association rules between couples of (MH/SH) are more precise than association rules between main headings, and between main headings and subheadings since a subheading specifies a particular aspect of a main heading. With the same thresholds as in cases 1 and 2, the number of rules is 2,565 (648 exact rules; 1,917 approximate rules).

The extracted association rules in the precedent cases are related to the medical domain. To obtain more precise rules we performed experiments on categorized documents according to groups of users: students in medicine, health professionals, and general public to evaluate the influence of categorization on the generation of association rules.

5.4. Categorizing documents according to health information seekers

In CISMeF, mainly three types of health information seekers are categorized: professionals, students in medicine, patients and lay people. We consider three major resource types: guidelines*, education* and patients*. We also consider two kinds of itemsets: the set of major main headings I={MH*} and the set of major (main heading/subheading) pairs I={[MH/SH]*)}. The collection is detailed in Table 6.

Resource type	Documents	Items	Min	Max	Mean
Guidelines*	2,727	MH*	1	64	5.21
		MH/SH*	1	70	6.12
Patients*	3,272	MH*	0	25	1.63
		MH/SH*	0	30	1.82
Education*	3,610	MH*	0	25	2.22
		MH/SH*	0	34	2.73

Table 4. Description of the collections of documents.

For all contexts, the minimum support threshold was fixed to minsup=20 and the minimum confidence threshold was fixed to minconf=70% (Table 7). We obtained association rules between major main headings MH* in the first context where I={MH*} and between [MH/SH]* pairs for I={[MH/SH]*}. For the major resource types patients* and education* all association rules (100%) are between two MHs* and between [MH/SH]* i.e. one descriptor in the antecedent and one descriptor in the consequent. For the major resource type guidelines*, 24% of the rules are between more than two descriptors. The characteristics of documents may explain these results: average descriptors were from 1.63 to 2.22 for patients* and education* whereas they were from 5.21 to 6.12 for guidelines*.

Resource types	Item=MH*				Item=[MH/SH]*			
	Nb rules	ER	AR	Nb pairs	Nb rules	ER	AR	Nb pairs
Guidelines*	50	12 24%	38 76%	38 76%	39	8 20.51%	31 79.49%	35 76%
Patients*	20	9 45%	11 55%	20 100%	19	8 42.1%	11 57.9%	19 100%
Education*	23	6 26.09%	17 73.91%	23 100%	25	13 52%	12 48%	25 100%

Table 5. Number of rules, exact rules (ER), approximate rules (AR), and number of pairs.

5.4.1. Evaluation of the extracted knowledge

Not all of the association rules extracted were evaluated: according to the context extraction and the itemset I there are more or less association rules. The more the collection is specialized, and the itemset size is reduced, the less we have association rules to evaluate. As defined, an interesting association rule confirms or states a new hypothesis [38].

Here, we proposed to combine background domain knowledge with simple statistical measures used traditionally in association rules mining for evaluation. We considered several cases of interesting association rules according to relations between MeSH headings. As these relations are defined between two main headings and between two subheadings, we considered only the association rules between two elements. Hence, an interesting existing association rule could associate: a (in)direct son and its father (relation FS); two descriptors that belong to the same hierarchy (same (in)direct father) (relation BR); two descriptors with See Also relation (relation SA). These rules are automatically classified thanks to the MeSH structure. The other rules that satisfy the minsup and minconf are then considered as «new» interesting association rules.

Exact association rules, except for collection patients*, are mostly new interesting rules: from 62.5% to 87.4%. Therefore, existing rules are mainly from the patients* collection: 77.8% for MH* and 75% for MH/SH*. However, approximate rules, are mostly existing rules (Table 8). Subjective interest measures are based on expert knowledge about the data, *i.e.* that of physicians and medical librarians in this context. New interesting rules for the contexts MH* and [MH/SH]* pairs are evaluated manually. 93.8% (resp. 84.8%) of the interesting new rules with conf=1 (resp. conf≥0.7) between major descriptors are validated.

Resource types	Items	Exact rules				Approximate rules			
		Existing knowledge			New	Existing knowledge			New
		FS	BR	SA		FS	BR	SA	
Guidelines*	MH*	-	-	4 33.3%	8 66.7%	2 5.3%	7 18.4%	10 26.3%	12 31.6%
	MH/SH*	1 12.5%	1 12.5%	1 12.5%	5 62.5%	3 9.7%	3 9.7%	9 29%	13 42%
Patients*	MH*	-	5 55.6%	2 22.2%	2 22.2%	2 18.2%	2 18.2%	4 36.3%	3 27.3%
	MH/SH*	-	5 62.5%	1 12.5%	2 25%	2 18.2%	2 18.2%	3 27.3%	7 36.3%
Education*	MH*	1 16.7%	1 16.7%	-	4 66.6%	2 11.8%	6 35.3%	3 17.6%	6 35.3%
	MH/SH*	1 7.7%	-	1 7.7%	11 87.4%	2 16.8%	3 25%	2 16.8%	5 41.4%

Table 6. Association rules evaluation according to the MeSH structure

5.5. Knowledge-based query expansion

Our objective is to re-use the numerous association rules that we extracted from the CISMeF database into the information-retrieval process by query expansion. We use Interactive Query Expansion. For example, the association rule *breast cancer → mammography* is extracted from the corpus because the keywords *breast cancer* and *mammography* are frequently used together to index the documents. This association rule is as a "new" one because it doesn't exist in the domain knowledge which is, in our case, the MeSH thesaurus. When applying the association rule *breast cancer → mammography* on a query containing the term *breast cancer*, an interactive query expansion proposes to the user e-health documents related to *mammography* to complete the search. In medicine and health-related information, [40] have already investigated an efficient algorithm for association rule mining using the MeSH thesaurus. They adopted a MeSH-indexed representation of MEDLINE records, but the evaluation of the interest of the mined associations with respect to the task of PubMed retrieval improvement was not considered by the authors. In [41] many other works on information retrieval and query expansion in the biomedical domain are also presented. Methods to perform query expansion with promising results involve mining user logs [41] and constructing user profiles. And another study on logs in PubMed for searching biomedical and life-science literature online has been performed by [43].

In the literature, a number of methods for performing query expansion have been developed. The solutions given are based mainly on two approaches. The first is the

augmentation of query terms to improve the retrieval process without user intervention. The second is the suggestion of new terms to the user which can to be added to the original query to guide the search towards a more specific document space. The first case is called automatic query expansion whereas the second case is called semi-automatic query-expansion. In [44], the authors tried to evaluate and compare the efficiency of the two methods. Despite the fact that their experiments were based on simulations and not on real human users in most of the cases, the results of the experiments showed that the interactive query expansion method gave more control to the searcher who knows her utility better than any automated system. Researchers also turned to methods such as lexical co-occurrence [45]. Lexical co-occurrence is the process of developing relationships between words based upon their co-occurrence in documents. The similarity of the method we have proposed here with lexical co-occurrence is that the source, which provides the candidate terms for expansion, is the set of the retrieved documents as opposed to some knowledge structure as in thesaurus-based approaches. As a consequence, if the user chooses terms that do not yield results from the expected domain, the terms suggested by the query-expansion algorithm are unlikely to be helpful to the user. A solution may be a simple spell-checker.

5.6. Evaluating query expansion based on association rules

Many ways of navigation and information retrieval are possible in the catalogue. The most used is the simple search (free text interface). As stated in the section 2, it is based on the subsumption relationships. A query (a word or an expression) can be matched with an existing concept. In this case, the result of the query is the union of the resources that are indexed by the concept, and the resources that are indexed by the concepts it subsumes, directly or indirectly, in all of the hierarchies it belongs to. The co-occurrence tools developed for information retrieval bring the terms which frequently appear in the same documents closer together. These terms thus have a semantic proximity. This technique was used very early to allow query expansion. By analogy, association rules may be exploited in a search engine by carrying out an interactive query expansion. This helps the user to formulate his query by using the result of a query to reformulate, filter and re-orientate the query by exploiting the terms related to his query terms. In fact, the user can select suggested terms sets to add them to his initial query. It is useful in the case of non-precise information needs. IQE requires user implication. We developed a web-based evaluation tool of the IQE used by a set of 500 users which are subscribers of the weekly letter “What’s new” of CISMeF. 20 queries, and for each one a set of medical terms derived from the extracted association rules were proposed. The evaluation was performed thanks to a Likert scale. The results (76% of the users were satisfied by the propositions) demonstrate the usefulness of this approach. An expanded query by association rules contains more related terms. By using the vectorial model, for example, more documents will be located and this treatment increases recall. In addition, association rules are indication on the possible definition of a term or its context.

6. Conclusions

We have presented in this chapter useful methods to help health information seekers to find resources on the Internet which is the most popular way used nowadays. The experiences were carried out on the CISMeF catalogue in French, but are reproducible for other e-health applications in other languages. These methods include simple ones such as heuristics and spell-checking, and more sophisticated ones such as knowledge extraction from e-health documents.

Author details

Lina F. Soualmia, Badisse Dahamna and Stéfan J. Darmoni
CISMeF & TIBS-LITIS EA 4108, Rouen University &Hospital, France

7. References

[1] Hou J, Zhang Y. Effectively finding relevant web pages from linkage information. IEEE Transactional Knowledge Data Engineering 2003; 15(4), 940–951.

[2] Liu, B. Web data mining: exploring hyperlinks, contents and usage data. Springer. 2007

[3] Keselman A, Browne AC, Kaufman DR. Consumer health information seeking as hypothesis testing. Journal of American Medical Informatics Association 2008; 51(4):484–495.

[4] Koch T. Quality-controlled subject gateways: definitions, typologies, empirical overview. Online Information Review 2000; 24(1):24–34.

[5] Abad Garcia F. A comparative study of six European databases of medically-oriented web resources. Journal of the Medical Library Association 2005; 93(4):467–479.

[6] Douyère M, Soualmia LF, Rogozan A, Dahamna B, Leroy JP, Thirion B, Darmoni SJ. Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. Health Information Libraries Journal 2004; 21(4):253–261.

[7] Baker T. A grammar of Dublin Core. D-Lib Magazine 2000; 6(10).

[8] Nelson SJ, Johnson WD, Humphreys BL. Relationships in Medical Subject Heading. In: Relationships in the Organization of Knowledge, 2001, eds. Kluwer Academic Publishers, pp. 171–184.

[9] Soualmia LF, Darmoni SJ. Combining Different Standards and Different Approaches for Health Information Retrieval in a Quality-controlled Gateway. International Journal of Medical Informatics 2005; N°74; vol (2-4); pp. 141–150.

[10] McCray AT, Ide NC, Loane RR, Tse T. Strategies for supporting consumer health information seeking. Proceedings of the 11th World Congress on Health Informatics, Medinfo 2004; pp.1152–1156.

[11] Grannis SJ, Overhag MJ, Mc Donald C: Real world performance of approximate string comparators for use in patient matching. Studies in Health Technology and Informatics 2004; 107:43–47.
[12] Levenshtein V. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Dokl 1966; 10:707–710.
[13] Yarkoni T, Balota D, Yap M. Moving beyond Coltheart's N: a new measure of orthographic similarity. Psychonomic Bulletin & Review 2008; 15(5):971–979.
[14] Soualmia LF. Towards intelligent information retrieval with query expansion knowledge-based methods. PhD thesis 2004; INSA Rouen.
[15] Hodge VJ, Austin J. A comparison of a novel neural spell checker and standard spell checking algorithms. Pattern Recognition 2002, 11(35):2571–2580.
[16] Damereau FJ. A technique for computer detection and correction of spelling errors. In Communication of the ACM 1964; 7(3):171–177.
[17] Peterson LJ. A note on undetected typing errors. Communications of ACM 1986. 29(7):633–637.
[18] Kuckich K. Techniques for automatically correcting words in text. ACM Comput Surv 1992, 24(4):377-439.
[19] Kernigham M et al. A spelling correction program based on noisy channel model. In proceedings of conference on COmputational LINGuistics, 1990. vol. 2.
[20] Brill E, Moore RC. An improved error model for noisy channel spelling correction. In proceedings of the Association for Comput. Linguistics 2000; 286–293.
[21] Toutanova K, Moore RC. Pronunciation Modeling for Improved Spelling Correction. In proceedings of the Association for Comput. Linguistics 2002; 141–151.
[22] Boyer C, Baujard V, Griesser V, Scherrer JR. HONselect: a multilingual and intelligent search tool integrating heterogeneous web resources. International Journal of Medical Informatics 2001, 64(2–3):253–258.
[23] Crowell J, Long Ngo QNG, Lacroix E. A frequency-based technique to improve the spelling suggestion rank in medical queries. Journal of the American Medical Informatics Association 2004, 11(3):179–185.
[24] Peters L, Kapunsik-Uner JE, Nguyen T, Bodenreider O. An approximate matching method for clinical drug name. AMIA Annual Symposium 2011, in press.
[25] Wilbur JW, Kim W, Xie N. Spelling correction in the PubMed search engine. Information retrieval 2006. 9:543–564.
[26] Stoilos G, Stamou G, Kollias S. A string metric for ontology alignment. In Proceedings of the International Semantic Web Conference, 2005; 624–637.
[27] Yujian L, Bo L. A normalized Levenshtein distance metric. IEEE Transactions on Pattern Analysis and Machine Intelligence 2007, 29(6):1091–1095.
[28] Winkler W. The state record linkage and current research problems. Technical report: Statistics of Income Division, Internal Revenue Service Publication 1999.

[29] Moalla Z, Soualmia LF, Prieur-Gaston E, Lecroq T & Darmoni SJ. Spell-checking queries by combining Levenshtein and Stoilos distances. Proceedings of Network Tools and Applications in Biology 2011, online.

[30] Wagner MM. An Automatic Indexing Method for Medical Documents. Symposium on Computer Application in Medical Care. 1991:1011–1017.

[31] Grosjean J, Merabti T, Dahamna B, Kergourlay I, Thirion B, Soualmia LF & Darmoni SJ. Health Multi-Terminology Portal: a semantics added-value for patient safety. Studies in Health Technology and Informatics 2011, 166:129-138.

[32] Soualmia LF, Griffon N, Grosjean J & Darmoni SJ. Improving Information Retrieval by Meta-modelling Medical Terminologies. Proceedings of 13th Conference on Artificial Intelligence in MEdicine 2011, 6747:215–219.

[33] Gehanno JF, Thirion B, Darmoni SJ. Evaluation of Meta-Concepts for Information Retrieval in a Quality-Controlled Health Gateway. In: Proceedings of the American Medical Informatics Association symposium 2007; pp. 269–273.

[34] Trombert-Paviot B, Rodrigues JM, Rogers JE, Baud R, van der Haring E, Rassinoux AM, Abrial V, Clavel L, Idir H. GALEN: a Third Generation Terminology Tool to Support a Multipurpose National Coding System for Surgical Procedures. In International Journal of Medical Informatics 2000; 58-59: 71–85.

[35] Spink A, Wolfram D, Jansen BJ & Saracevic T. Searching the web: the public and their queries. Journal of the American Society Information Science Technology 2002, 52(3):226–234.

[36] Agrawal R, Imielinski T & Swami AN. Mining association rules between sets of items in large databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data 2003; 207–216.

[37] Agrawal R. & Srikant R. Fast algorithms for mining association rules in large databases. In Proceedings of the VLDB Conference 1994; pp. 478–499.

[38] Pasquier N, Taouil R, Bastide Y, Stumme G & Lakhal L. Generating a condensed representation of association rules. Journal of Intelligent Information Systems 2005, 24(1):29–60.

[39] Fayyad UM, Piatetsky-Shapiro GP, Smyth P & Uthurusamy R. Advances in Knowledge Discovery and Data Mining 1996. American Association of Artificial Intelligence.

[40] Kahng J, Liao WHK & McLeod D. Mining generalized term associations: count propagation algorithm. In Proceedings of the KDD workshop 1997, pp.203–206.

[41] Prince V & Roche M. Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration 2009. IGI Global.

[42] Cui H, Wen JR & Ma WY. Query expansion and classification by mining user logs. Knowledge and Data Engineering 2003, 15 (4):829–839.

[43] Lu Z. & Wilbur WJ. Improving Accuracy for identifying related PubMed queries by an integrated approach. Journal of Biomedical Informatics 2009; 42:831–838.

[44] Ruthven I. Reexamining the potential effectiveness of interactive query expansion. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 2003; pp. 213–220.

[45] Vechtomova O, Robertson S. & Jones S. Query expansion with long-span collocates. Information Retrieval 2003, 6(2), 251–273.

IntechOpen

IntechOpen