International Journal of Medical Informatics (2006) **75**, 73—85

ELSEVIER

International Journal of Medical Informatics

www.intl.elsevierhealth.com/journals/ijmi

Health search engine with e-document analysis for reliable search results

Arnaud Gaudinat[a,*], Patrick Ruch[b], Michel Joubert[c], Philippe Uziel[d], Anne Strauss[h], Michèle Thonnet[e], Robert Baud[b], Stéphane Spahni[f], Patrick Weber[f], Juan Bonal[g], Celia Boyer[a], Marius Fieschi[c], Antoine Geissbuhler[a,b]

[a] *Health on the Net Foundation, Geneva, Switzerland*
[b] *SIM, University Hospital of Geneva, Switzerland*
[c] *LERTIM, Timone Hospital, Marseille, France*
[d] *XR partner, Paris, France*
[e] *MISS, Paris, France*
[f] *NICE Computing, Lausanne, Switzerland*
[g] *THALES Information System, Paris, France*
[h] *Institut de Recherche pour le Développement (IRD), Paris, France*

KEYWORDS
Web Search engine;
Trustworthy information;
eHealth

Summary
Objective: After a review of the existing practical solution available to the citizen to retrieve eHealth document, the paper describes an original specialized search engine WRAPIN.
Method: WRAPIN uses advanced cross lingual information retrieval technologies to check information quality by synthesizing medical concepts, conclusions and references contained in the health literature, to identify accurate, relevant sources. Thanks to MeSH terminology [1] (Medical Subject Headings from the U.S. National Library of Medicine) and advanced approaches such as conclusion extraction from structured document, reformulation of the query, WRAPIN offers to the user a privileged access to navigate through multilingual documents without language or medical prerequisites.
Results: The results of an evaluation conducted on the WRAPIN prototype show that results of the WRAPIN search engine are perceived as informative 65% (59% for a general-purpose search engine), reliable and trustworthy 72% (41% for the other engine) by users. But it leaves room for improvement such as the increase of database coverage, the explanation of the original functionalities and an audience adaptability.

* Corresponding author.
E-mail address: arnaud.gaudinat@healthonnet.org (A. Gaudinat).

1386-5056/$ — see front matter
doi:10.1016/j.ijmedinf.2005.11.002

Conclusion: Thanks to evaluation outcomes, WRAPIN is now in exploitation on the HON web site (http://www.healthonnet.org), free of charge. Intended to the citizen it is a good alternative to general-purpose search engines when the user looks up trustworthy health and medical information or wants to check automatically a doubtful content of a Web page.

1. Introduction

Since the Big Bang of the Worldwide Web over a decade ago, emanating from CERN, the hypertext universe has been in constant expansion and accelerating. Ubiquitous and pervasive yet easily accessible, the Web is today's oracle, with an instant answer to any type of question. Like the all-encompassing library of Babel in the story by Borges [2], the Web can satisfy our every curiosity, yet it is a simple mirror of human thought, mixing the exalted with the vile. The Web, with a thousand responses to every question, offers no clear answer to the searcher, whose quest leads ever deeper into the vastness of undifferentiated knowledge. The Web suffers from the overabundance of information, and the highly variable quality of its content. On trivial matters, a myriad of sources answer with a common voice, but it is more difficult to obtain an authoritative response to vital questions in the field of health, where charlatans have found a comfortable and lucrative home on the Internet.

1.1. State of the art

Currently, most citizens and patients use general purpose search engines (80% according to Jansen [3] such as Google 48% or Yahoo 21.2% when searching on the Web, for April 2005 according to Nielsen/Netrating [4]). According to Jansen [3], health and science occupy the 4th and 6th places, respectively, for the years studied (1997—2002 for different search engines), showing the importance of the health domain for the Internet citizen.

Other popular search tools include thematic directories such as Yahoo or the Open Directory Project (DMOZ). Specialized in the health domain are CISMeF [5] (only available in French) and HONselect [6] (in five languages) which present medical information arranged under the Medical Subject Headings thesaurus (MeSH) of the U.S. National Library of Medicine. HONselect also offers advanced multilingual features to facilitate comprehension of web pages in languages other than those of the user.

The search engines have dealt remarkably well with the ever-increasing volume of information, with over 8 billion pages now indexed by Google. Less certain is the ability of the general search engines to produce quality results as the database size increases. The spectacular success of Google [7] is probably due to its patented ''PageRank'' algorithm Brin and co-workers [8] (or other algorithms based on page popularity), based on the notion that a hyperlink from document A to document B implies that the authors of document A consider document B to be of value. It would be reasonable to assume that a page's popularity would be correlated to the quality of its content. A popular site could ill afford to spread bad information, as shown by studies such as Amento et al. [9], and especially in the medical field by Borges et al. [10]. Are we therefore to conclude that search engine results are indicative of quality? What about quality pages that lack 'popularity', such as new pages or those whose owner has not undertaken the promotional efforts often needed to obtain backlinks from reputed websites? Only an in-depth study would reveal with certitude whether a relationship exists between the number of inbound links and the quality of web page content as judged by human experts.

Other practical approaches have been put forth based on adherence to standard or selection by a third party. These include HON [11], AFGIS,[1] WMA [12], and URAC [13]. The HONcode of Health on the Net Foundation [14,15] is the unique example to have been deployed on a large scale, with over 5000 sites in 29 languages enrolled in a voluntary accreditation program. The accreditation process is initiated by the site operator, who prepares the site for review by a HON medical reviewer who checks for compliance with eight principles. Accreditation is free and remains in force as long as the site continues to pass an annual review. CISMeF and MEDLINEplus [16] are interesting initiatives whereby librarians select quality resources, respectively, for scholar content and limited to French-language sites, and intended for patients

[1] AFGIS: http://www.afgis.de/.

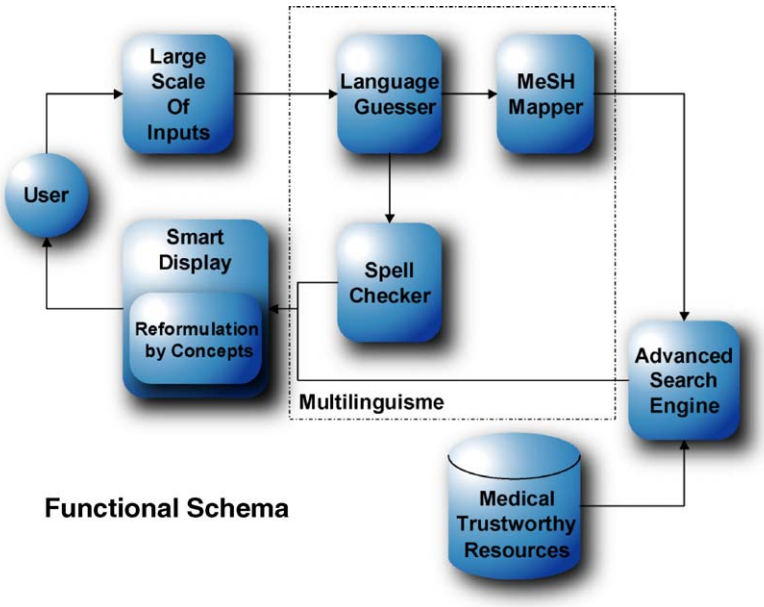

Fig. 1 Functional schema of WRAPIN as a guide of this article.

covering English and Spanish languages web sites. Also, the MedCertain project has produced the HIDDEL [17] markup language allowing formalization of a number of descriptive criteria for web page content in great detail. The system may, however, be difficult to apply on a large scale due to the need for web publishers to add the complex markup to each page. Human review remains necessary with HIDDEL, to control misuse (or abuse) of the language.

Numerous actors have thus contributed to the effort to limit the potential for harm resulting from poor quality online health information. Various approaches have been attempted but none offers a definitive solution, especially with regard to inexperienced Internet users, persons with a low degree of health literacy or those whose search is motivated by a medical crisis. Recent advances in search ranking technology attempt to take account of topicalization, or theming, see Haveliwala [18]. This narrows the search to pages for a relevant topic, but still does not deal with the quality of the content itself. General purpose search engines appear to be making little progress toward the analysis of information quality. In the critical field of health information, it is preferable to guide users to restricted domain search tools, which, while containing fewer resources, are more likely to provide reliable, relevant results. Our experience has shown that quality information is best delivered by user-centric search tools that favor resources created to benefit the end users rather than the information providers.

2. Materials and methods

In this section, we survey the set of methods and resources developed or simply used in the WRAPIN project.

2.1. Introduction to WRAPIN

To create an efficient search tool incorporating best practices for quality medical information, a group of medical informatics/Internet experts from HON, LERTIM, NICE and SIM proposed an innovative solution in the form of the EU-WRAPIN project (World Reliable Advice for Patients and Individuals). WRAPIN offers answers to many of the problems described above and was the fruit of multidisciplinary collaboration with experts from organizations including MISS and THALES IS. The main purpose of WRAPIN is to help users assess the credibility of online medical information, using a reference base constituted exclusively of trustworthy documents.

To accomplish this WRAPIN needed advances features which will be described in the following paragraphs. The simplified functional diagram below (Fig. 1) outlines the main WRAPIN functionalities and their interactions.

2.2. Medical trustworthy resources

The initial goal of WRAPIN was to make available, in the first place for patients, but also for professionals, a tool allowing the assessment of web content quality for the medical field. In a general

way, WRAPIN represents an alternative to existing search engines, as it combines the best medical Web pages with other 'hidden' online documents that are not referenced by other search engines. A sample of information resources is presented, giving the user an appreciation of the depth and breadth of scientific debate, agreement and controversy for a given subject, and a better understanding of the relationships linking diverse ideas and actors. The choice of reference sources is critical for WRAPIN and careful selection is required. Currently, eight sources are used by the system: PubMed of the U.S. National Library of Medicine (NLM), which counts over 15 million citations, ClinicalTrials.gov, also from NLM with the results of more than 20,000 studies, MedHunt (HON), with some 70,000 selected web pages, HONcodeHunt made up of over 110,000 trustworthy webpages from the HONcode-compliant sites, BookShelf from NLM with over 24,000 pages from digitized medical reference texts; HONNews with some 13,000 medical news items from HealthDay and HON; FDA (U.S. Food and Drug Administration) with a selection of more than 1000 relevant pages from the US FDA; OESO, from the OESO Foundation, a collection of 1153 scientific articles in the form of questions and answers from the field of esophageal disorders; and UroFrance, a database of over 2500 French-language articles on urology. By indexing these resources, WRAPIN covers an important range of medical subjects, merging scientific/technical documents with more accessible information destined for general readers including newly diagnosed patients and those with chronic illnesses as well as health professionals.

2.3. A large scale of inputs

The greatest innovative feature of WRAPIN is its ability to handle different types of query, especially entire web pages (specified by URL) as a query. Whereas, certain search tools can find related pages for a given page using a vector approach, WRAPIN analyses a page for the most important medical terms, performing a frequencies analysis on MeSH terms found on the page. WRAPIN identifies keywords which are then used for: (1) weighted queries to its indexes; (2) translations into languages other than that of the initial query; and (3) serve to highlight the most important medical concepts dealt with by the document. This approach is also applied to long texts or natural language expressions that are submitted by users as queries. This functionality opens up new possibilities for online information searching compared with existing search tools (including those in the medical field) which limit queries to just a few words. Theoretically — and practically — a voluminous document such as an academic thesis could be used to query WRAPIN, if computational time and resources are available. Applications such as bibliographic research could be reasonably envisaged in such cases. Alternatively, a short query input is handled by standard means, with the addition language translations and weighting by frequency of medical terms. The figure below presents the main interface of WRAPIN, showing the first field for entry of a URL and the second permitting entry of an arbitrary block of text (Fig. 2).

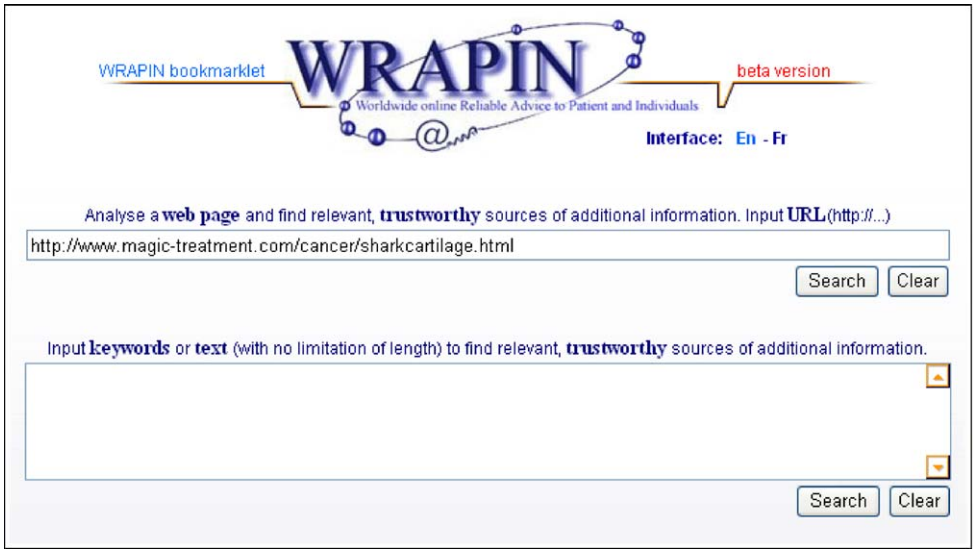

Fig. 2 WRAPIN interface (URL, short query or text with no limitation of length).

2.4. MeSH mapper

The MeSH (Medical Subject Headings) of the U.S. National Library of Medicine is a thesaurus [1], a hierarchically arranged terminological resource for the medical domain containing 22,997 descriptors. It was created to meet the need for indexing of medical literature. MEDLINE has proved its usefulness over many years. The specialized librarians of MEDLINE have carefully selected MeSH terms for over 15 million scientific articles, some of which date back to the 1950s. The use of a thesaurus for indexing is not a new idea [19] and has been widely applied for medical literature [20]. The MeSH remains a precious resource when it comes to manipulating multilingual medical text data, or when performing queries expansion. After HONSelect, the first multilingual repertory based on the MeSH hierarchy, and CISMeF, catalog of French-language medical resources based on the MeSH, WRAPIN relies on automatic categorization of queries and documents based on this nomenclature.

A first challenge is to efficiently extract MeSH terms from the analyzed documents. A great amount of research on concept recognition in medical text has already been done [19—22]. The best known is probably the indexing initiative [21] from the U.S. National Library of Medicine that resulted in the well-known MetaMap system [22] and the related citation algorithm [23], where the goal is to help or replace the human annotator of MEDLINE. For WRAPIN, we investigated a non-supervised approach based on a space vector model, which provided results as good as those that can be obtained with MetaMap for the MeSH term extraction task [24]. Recently, this WRAPIN module (called, HonMeSHmapper system) participated in an evaluation session with other French Language mapper systems (conducted by Névéol et al. [25]), in which the results showed that HONMeSHMapper achieved the best overall *F*-measure.

The mapping of MeSH terms is crucial within WRAPIN, as it is used throughout the system, for queries as well as for documents. Mapping serves to categorize text using the most representative MeSH headings, which are needed at the time of indexing, searching, translation, scoring, in query reformulation, and for formatting of the results page.

Table 1 presents an example of MeSH mapping (or research for key concepts) following analysis of a web page. The page in question has as its purpose the sale of shark cartilage for the treatment of cancer. WRAPIN attempts to identify key concepts from the page in order to create a synthetic query to trustworthy databases. These key concepts are listed in order of decreasing importance. The first column shows how the word (i.e. Entry Term or MeSH heading in case of a MeSH term) appears within the text; a frequency is associated with each word. Here however, the concept 'shark cartilage' is not a MeSH term, but because of its frequent occurrence in the page is considered as relevant (WRAPIN is capable of identifying key terms composed of up to three words). In the third column, the terms are grouped by key concepts (e.g. MeSH heading). To the right of each key concept, the type (MeSH or other) is listed, followed by the cumulative frequency of occurrence for the different forms of each concept and, finally, a score based on the cumulative frequency and the

Table 1 MeSH mapping results for the URL ''http://www.discount-vitamins-herbs.net/shark-cartilage.htm''

In text	Frequency	Key concept	Type	4.1.1.1.1 Total	4.1.1.1.2 Score
Cartilage	306	Cartilage	MeSH	306	0.047
Cancer	102	Neoplasms	MeSH	177	0.027
Tumor	36				
Tumors	36				
Cancers	3				
Cells	111	Cells	MeSH	165	0.025
Cell	54				
Treatment	96	Therapeutics	MeSH	153	0.023
Treated	30				
Treatments	12				
Therapeutic	9				
Treat	6				
Shark cartilage	146	Shark cartilage	Other	146	0.022
Angiogenesis inhibitors	42	Angiogenesis inhibitors	MeSH	42	0.012

type (medical or non-medical) and its inverse frequency (this method favors MeSH terms).

The work done during the WRAPIN project shows that the UMLS (Unified Medical Language System of the U.S. National Library of Medicine) knowledge sources may contribute to a better indexing of medical documents by the use of MeSH terms [26]. Special attention has been focused on a (very large) piece of knowledge contained in the UMLS knowledge sources: co-occurrences between major MeSH terms in the Medline literature [27]. Previous work within the ARIANE project demonstrated a way to translate semantic relationships between concepts into MeSH sub-headings [28]. The reverse is done in WRAPIN: translation of sub-headings associated with several terms by the UMLS, according to co-occurrence frequencies, into relationships between the concepts represented by the terms. The aim is to propose to the indexer possible semantic associations between concepts it has recognized in analyzed texts. When associations are validated according to this knowledge database, the indexer is able to refine its results.

UMLS is a project of the U.S. National Library of Medicine. UMLS has two main components: the Metathesaurus and the Semantic Network. The Metathesaurus contains not only MeSH but also the most useful medical nomenclatures. The core concepts of the Metathesaurus are connected to generic types of concepts in the Semantic Network. These types of concepts are interconnected by semantic relationships [29]. The data structure of the Metathesaurus is based on hierarchies and associations. The association relationship links a given term to related terms and to a preferred term. The (pre-)order relationship structures the preferred terms into more generic terms and more specific ones. This later relation divides the thesaurus into several so-called microthesauri, according to a local specificity. For example, the term Coronary Arteriosclerosis appears twice in the Metathesaurus: firstly, as a process involved in coronary diseases viewed as heart diseases, and secondly, as an arteriosclerosis localized into the coronary arteries and causing arterial occlusive diseases. The presence of micro-thesauri translates the various contexts from which a same medical concept can be viewed and, thus, the complexity of the medical domain.

The Semantic Network associates types of medical concepts with semantic relationships. The types of concepts are organized in a hierarchy where, for instance, Physiologic Function and Pathologic Function are children of Biologic Function, and Disease or Syndrome is a child of Pathologic Function. There are about 30 different semantic relationships. Among them, for instance, Diagnoses applies on the two types, Diagnostic Procedure and Pathologic Function, Treats applies on Therapeutic or Preventive Procedure and Pathologic Function. These semantic relationships are defined at such a general level that is not always possible to map a type onto its linked concepts by an automatic pertinent inheritance of the meaning that the relationships convey: it is obvious that every diagnostic procedure cannot be used to diagnose every pathologic function. Nevertheless, since a Coronarography is linked to the type Diagnostic Procedure, it can be used to diagnose a Coronary Arteriosclerosis that is linked to the type Disease or Syndrome and thus to the type Pathologic Function. The Semantic Network provides today a framework for biomedical concepts isolated into the Metathesaurus that can be considered as an operational ontology.

Our working hypothesis, which is then applied in treatments, is that significant associations between concepts are those for which there exist semantic relationships in the literature materialized by co-occurrences in the above UMLS knowledge source. The exploitation of this hypothesis is done on extracted MeSH terms and produces a ranking list of candidate terms, which are those terms that best characterize a document. Two steps make up the process: (1) for each pair of terms present in the list of terms, a cumulative weight of their relationships is computed, and (2) the weight affected to each term is then computed according to all the pairs in which it is present. This approach has been successfully experimented [30].

2.5. Advanced search engine

The search engine is the hub of an application such as WRAPIN. Pioneers like Salton and McGill [31] knew how to use the nature of information itself to create efficient models and technologies. Conferences such as TREC [32] have helped further our knowledge of information searching in general and online search engines in particular.

Within WRAPIN exist various types of documents; from MEDLINE citations in XML format to HTML pages created by practicing physicians, there are many types in between, more or less structured, all of which need to be handled by one system. The goal is to get at the essence of each document, by the use of MeSH terms for indexing for MEDLINE documents, analysis of formatting markup for HTML, to identify key elements in the text. For reasons of efficiency and coherence in the calculation of scores, all of the trustworthy medical resources cited in this paper were indexed locally. Information searching with WRAPIN is based on

evidence developed over the past decades in the field of classic information searching, as well as more recent research into web-based information searching, and takes into account the specificities of the medical domain, making use of the numerous specialized resources available (terminological, semantic, etc.). Searching within WRAPIN use principles whereby: (1) the frequency of terms in each document as well as the inverse frequency of documents for each term (model known as TF-IDF); (2) MeSH are boosted; (3) synonyms of found MeSH terms are used; and (4) the spatial proximity of keywords is taken into account. The TF-IDF is applied to the title, content and URL of the document.

2.6. Smart display results

Presentation of results is of the greatest importance for search engines, many of which have chosen a simple, sober design, while others have opted for a complex interface which, while intriguing, may be daunting to new users. The inspiration of the WRAPIN interface comes from the simplicity of classic web search tools, with added functionality appropriate to the medical domain. The most interesting functionality is without a doubt the automatic identification of the conclusion, as proposed by Ruch et al. [33] for implicitly structured documents, such as MEDLINE abstracts, to complement the classic KWiC (KeyWord in Context) extraction. The basic hypothesis behind this is that the conclusion of a scientific article contains the most relevant information for the searcher.

2.6.1. Key sentences with latent argumentative structuring

Key word assignment has been largely used in MEDLINE to provide an indicative ''gist'' of the content of articles. Abstracts are also used for this purpose. However, with usually more than 300 words, abstracts can still be regarded as long documents; therefore we designed a system to select a unique key sentence from a MEDLINE abstract. Following recent developments in information retrieval [34] and machine learning [35], which show that conclusions are the most content-bearing sentences to perform related articles search and index pruning tasks in MEDLINE, we assume that conclusion sentences would be good candidates for such key sentences in scientific texts. Selecting argumentative contents is formally a classification task: for each piece of text the system will have to decide whether it is a relevant conclusion or not. In text classification tasks, two types of strategies are competing: expert-driven and data-driven approaches. While the former, which rely on a domain expert, are often time and labour-intensive, the latter are directly dependent on the availability of large training sets. Fortunately, training data for our task can be acquired in a cheap manner. Most abstracts in MEDLINE are unstructured (i.e. provided without explicit argumentative markers, such as METHODS, PURPOSES, ...), but fortunately, a significant fraction of these abstracts contain explicit argumentative markers. Using PubMed and its Boolean query interface, we collected a set of 12,000 MEDLINE citations containing strings such as ''PURPOSE:'', ''METHODS:'', ''RESULTS'', ''CONCLUSION:'' (cf. Fig. 3).

We use a Bayesian classifier which has the advantage of showing linear complexity [36], while most other top performing algorithms tend to have a quadratic complexity; therefore they are often more adapted for rapid application developments and exploratory studies [37,38]. Three types of features are linearly combined to get a final probability ranking per class: stems; stem bigrams and stem trigrams. This approach has been evaluated and finally for the CONCLUSION class, the *F*-score (i.e. the harmonic mean, with recall and precision having the same importance; cf [39]) reaches 84%. While recall shows excellent effectiveness, precision could still be improved: conclusion segments are well classified, but some non-conclusion

INTRODUCTION: Chromophobe renal cell carcinoma (CCRC) comprises 5% of neoplasms of renal tubular epithelium. CCRC may have a slightly better prognosis than clear cell carcinoma, but outcome data are limited. ***PURPOSE:*** In this study, we analyzed 250 renal cell carcinomas to a) determine frequency of CCRC at our Hospital and b) analyze clinical and pathologic features of CCRCs. ***METHODS:*** A total of 250 renal carcinomas were analyzed between March 1990 and March 1999. Tumors were classified according to wellestablished histologic criteria to determine stage of disease; the system proposed by Robson was used. ***RESULTS:*** Of 250 renal cell carcinomas analyzed, 36 were classified as chromophobe renal cell carcinoma, representing 14% of the group studied. The tumors had an average diameter of 14 cm. Robson staging was possible in all cases, and 10 patients were stage 1) 11 stage II; 10 stage III, and five stage IV. The average follow-up period was 4 years and 18 (53%) patients were alive without disease. ***CONCLUSION:*** The highly favorable pathologic stage (RI-RII, 58%) and the fact that the majority of patients were alive and disease-free suggested a more favorable prognosis for this type of renal cell carcinoma.

Fig. 3 Example of explicitly structured abstracts in MEDLINE.

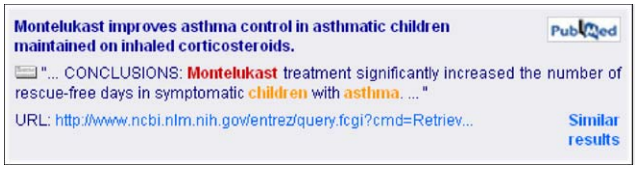

Fig. 4 Automatic conclusion detection for the query ''Montelukast for children with asthma''.

sentences are classified as conclusion (false positives). This is problematic for RESULTS segments, which is found ill-defined by the classifier, but looking at the corpus, the distinction between RESULTS and CONCLUSION appears questionable, so that merging these two classes could be both beneficial and legitimate. So, Naive Bayes classifiers provide an adapted framework to perform argumentative classification and outperforming expert-driven approaches.

Fig. 4 presents an example of a conclusion returned on a results page from WRAPIN. This example perfectly illustrates the usefulness of this functionality for the query 'Montelukast for children with asthma' where the conclusion provides a clear answer to the user's query. However, it happens that the conclusion, while providing an argumentative summary, is not directly related to the informative content of the query. Therefore the selected passage is displayed to the user only when it shares a minimal lexical similarity with the query. Otherwise, the classic KeyWord in Context (KWiC) display is often a more appropriate choice for category-driven passage extraction.

2.6.2. Advanced Keyword In Context system

In the case of non-structured documents such as web pages, a classical solution was applied. The use of a KWiC algorithm offers users an approximate synthesis of the document. In our case, it attempts to find the segments that maximize the number of medical terms and other keywords (a kind of ''keyword diversity'', including synonyms). For this search for segments, each term is weighted according to its weight in the initial query, where medical terms are favored (Fig. 5).

In both types of resume, each medical term (MeSH) is displayed in orange (with rapid access to its definition, for synonyms as well) and in red for non-medical keywords.

2.7. Reformulation by concepts

According to Hearst [40] search engine users often have only a vague idea of how to express the object of their queries. A great number of queries submitted to search engines contain only a word or two, according to Jansen and Spink [41], with the result that the multitude of documents returned may be relevant to the (vague) query, but useless to the searcher. In light of these considerations, an efficient search system should interact with the user, offering meaningful suggestions to add precision to the query. This would be especially valuable in the medical field where users may not have the requisite knowledge or vocabulary to specify the desired results. Web search engines now use clustering methods (the best know of these is probably Vivissimo [42]) based on automatic classification approaches that allow users to refine and specify queries according to subject classes found automatically in the documents returned for a query [43]. WRAPIN offers similar functionality based exclusively on the MeSH. Preprocessing uses the MeSH to categorize all documents prior to indexing, then perform a query and return the MeSH terms common to the first 10 documents and propose a list of terms related to: (1) the query (intrinsically) and (2) the documents for the query.

Fig. 5 Sample of WRAPIN results for the query ''fever children aspirin''.

Table 2 Reformulation suggestions for the query ''fever children aspirin''

Reye syndrome
Respiratory tract infections
Acetaminophen
Bacterial infections
Infection
Viruses
Pneumonia
Pharyngitis
Bronchitis chronic
Pharynx
Mucus
Common cold
Respiratory sounds

Table 2 presents a list of MeSH terms proposed by WRAPIN for the query ''fever children aspirin''. The first term returned, ''Reye syndrome'' is entirely relevant since the use of aspirin to treat fever in children can provoke this condition; as an alternate treatment one can use ''acetaminophen'', which is offered as a third MeSH term. A major difference of this system is that it does not restrict as a result of user interaction the user's search space, instead proposing concepts related to the original query, in order to refine it.

2.8. Multilingual approach, Spell checking and language guessing

Drawing on experience from HONSelect, WRAPIN uses the MeSH as departure point for multilingual functionality. Analysis of the query allows recognition of the most relevant MeSH terms in the query. These terms are then translated into the other languages for which MeSH has been made available within WRAPIN (English, German, French, Spanish and Portuguese). The advantage of using a thesaurus for the translation in the CLIR (Cross Language Information Retrieval) has been shown, notably by Eichmann [44]. This researcher uses the OHSUMED corpus, with queries translated manually into Spanish and French with the goal of finding the same documents in all three languages following automatic translation of the queries into English using UMLS (with various strategies) as the central terminology. Very good results have been obtained for Spanish with less favorable results for French, which he attributes to linguistic differences. In our case, we use a stemmer for French, synonyms, and the 2005 version of the MeSH, which undergoes annual refinements. In WRAPIN, four queries are performed in parallel in order to query the databases in the other languages.

Fig. 6 shows part of a page of results for the query ''fever children aspirin''. This query (translated) produced 84 documents in French, 46 in Spanish and 7 in German. Fig. 7 presents the same page with the MeSH terms translated into French. Note that the reformulation is now offered in French, since the query has been translated into French.

Prior to handling of the query, a language detector and spelling corrector are called. The latter is valuable for non-professional users who may approach the medical domain in an approximate way (for a discussion of the importance of spelling correction in information searching, see [45]). The correction tackles the form of a single suggestion, while the query continues to execute. This style of correction is also known by the name, ''Did You mean'', popularized by Google. In his evaluation of the correction tools used by NLM, Crowell et al. [46] reports that the tool GNU Aspell [47] gives better results than the tool GNU Gspell. Our corrector is based on Aspell and a classic DTW (better known for text under the name, Levenstein edit distance [48]) to select the most similar candidate, where medical terms are favored. Aspell has the advantage of a large panel of dictionaries (52 languages in the version we used). Medical terms were added using the MeSH in the different languages used in WRAPIN,

Fig. 6 Results for the query ''fever children aspirin''.

Fig. 7 Multilingual results for the query ''fever children aspirin''.

since Aspell allows creation of user-defined dictionaries. The language detector is a hybrid, combining a lexical approach based on a list of stopwords and the MeSH with an ngram approach (cf. [49]), also derived from the MeSH.

3. WRAPIN evaluation

The WRAPIN search engine in exploitation today is the result of the prototype tested and improved by the evaluation outcomes.

The evaluation presented here has been conducted on the prototype in order to assess the relevance of the WRAPIN concept and functionalities. The evaluation has been conducted at the end of the project (March 2004). The intent was to expose various citizens, patients and individuals' and medical professionals, to the concrete use of WRAPIN's functionalities and to analyze their perception in terms of ergonomic and perceived quality and usability of the replies given by WRAPIN for a set of health questions proposed. These health questions have been extracted from FAQ (Frequently Asked Questions) found on general public medical portals and disease specific sites dedicated to the citizens.

4. Results

The use of WRAPIN has, on the whole, been perceived as informative 65% (59% for the other engine), reliable and trustworthy 72% (41% for a general-purpose search engine) by users. Yet in a number of cases, but this was also true for the general-purpose search engine, the replies given by the system were irrelevant. The rate of irrelevance/impertinence has been scientifically reckoned, and finally reduced. The evaluation results show that the upfront lexical analysis of the queries, as well as the ergonomic of the reformulation left room for improvement. The results show that future developments should embody the capacity for WRAPIN to cover a larger number of specialized bases (as well as larger bases) and subjects. Better information about the capacities of WRAPIN is essential. Some functions not available with any other search engine, such as the URL evaluation need to be clearly explained to the users.

Concerning the man/machine interface, and its ergonomic in general, WRAPIN needs more improvements at three different levels: general layout of the pages, better management of colors, fonts, logos and pictograms, light on line help, and a little written documentation.

With a view to exploitation, the audience question remains central: WRAPIN has to be adapted to an audience of citizens, or to differentiate between a 'technical WRAPIN' turned towards health professionals, and a 'patients and individuals WRAPIN'. In order to take into account this result WRAPIN has introduced the ''stethoscope'' pictogram to designate those more technical databases.

In the view of the testers, WRAPIN will overperform other engines when the replies are pertinent. By providing a synthetic and reliable reply to a query, or an assessment of a document further to the submission of an URL, WRAPIN goes beyond the other engines, and brings a valuable tool to users seeking trustworthy medical and health care information.

As regards content, the evaluation has shown that users would tend to trust WRAPIN more than other engines, due to the fact that certified material is considered, against lower quality material, if not obnoxious material, possibly some times in other cases. Nevertheless, trustworthiness is fragile and needs to be constantly reinforced.

In the case of WRAPIN, trustworthiness can be broken down into at least three main components: algorithmic trustworthiness, data and information trustworthiness and organisation trustworthiness.

5. Discussion

WRAPIN offers in many different ways innovating functionalities related to the search of medical and health online information. WRAPIN specialized on the health domain and with its features and know-how tries to remedy to the gaps of the most popular Web search tools. These general-purpose search engines, not really adapted to the quality of information, benefit of great success because of the large coverage of indexing and the rapidity of the processing. If the algorithms such as PageRanking make it possible to integrate in an inherent way a concept of quality of information yet it makes it difficult to quantify it. Studies made on this subject are too approximate in order to reach any final conclusion especially in the health domain where skepticism is required.

The alternative suggested by WRAPIN is founded on sources of reliable information and on functionalities which are based on detailed terminological resources specifically related to the medical domain (not easily transposable to other domains because less studied). The reliable resources being fewer, the major disadvantage of the approach is without doubt the lack of coverage. But at a guess of the redundancy of information present on the Web and the increased number of pages of poor quality, let us guarantee, following the example of foodstuffs, that the users prefer a better quality to quantity.

The evaluation enabled us to check that our approach was useful and that it answers to the expectation of most users. Since this evaluation, WRAPIN has greatly evolved and has still got some ways to go. For resources reasons, the algorithm of the PageRanking type was not integrated in WRAPIN but its adaptation should make it possible to obtain better results. However, it is difficult to quantify in advance the interest of this method within WRAPIN because of the particularity of the sub-networks produced by quality resources.

The integration of new sources of information is relatively easy. With regard to the Web, the coverage of the tools such as MedHunt and HONcodeHunt should be improved soon while respecting at any time the quality of this information.

In addition to the ''relevance'' of information, the quality is indeed an important dimension that is often excluded from most Web search engines. One of the important points concerning quality, also ignored among Web tools, is certainly the freshness of the information (i.e. update date of the document). Even if the advanced options of search engines propose to search for a period of time, the feature is related to the date of the indexing date and not the date of update of the document. So, this functionality is useless. On the contrary reference sites such as PubMed, logically favor the latest articles for a given subject. In pursuing the same goal, HONs News allows an intelligent combination between the relevance of the query and its publication date. Why would relevant news from the point of view of the subject be useful if the information is obsolete? Or what is the interest to obtain the latest news if it does not correspond to its query? With regard to WRAPIN, resources such as MEDLINE or News authorize a search according to the date. Moreover, principle 4 of the HONcode—information must be documented: referenced and dated—intrinsically allows to guarantee that the updated information is at least available on the pages of accredited sites. WRAPIN is therefore not the ultimate solution to the existing problems of the health Web. But with its innovating and multilingual functionalities and its reliable resources, it is certainly the most powerful alternative in the health domain when it comes to navigate with trust on this extraordinary resource that is the World Wide Web.

We wish to dedicate this article to the memory of the Professor Jean-Raoul Scherrer, the initiator of this new generation of search engine.

Acknowledgements

The WRAPIN project, IST-2001-33260, has been supported by the European Commission and the ''Office Fédéral de l'Education et la Science'' (OFES, Switzerland). The authors wish to thank all the WRAPIN members as well as the WRAPIN tester with their precious contribution in this ambitious project.

References

[1] National Library of Medicine, Medical subject headings, http://www.nlm.nih.gov/mesh.
[2] J.L. Borges, The library of Babel, in: Labyrinths, Selected Stories, & Other Writings, New Directions Pub. Corp., New York, 1964.
[3] B.J. Jansen, A. Spink, How are we searching the World Wide Web? A comparison of nine search engine transaction logs, Information Processing and Management, 2004.
[4] Nielsen NetRatings Search Engine Ratings: http://searchenginewatch.com/reports/article.php/2156451 (June 2005).
[5] S.J. Darmoni, J.P. Leroy, F. Baudic, M. Douyère, J. Piot, B. Thirion, CISMeF: a structured health resource guide, Methods Inf. Med. 39 (1) (2000) 30.
[6] C. Boyer, V. Baujard, V. Griesser, J.R. Scherrer, Rela, Links HONselect: a multilingual and intelligent search tool inte-

grating heterogeneous web resources, Int. J. Med. Inform. 64 (2—3) (2001) 253—258.
[7] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, vol.3, in: Proceedings of the Seventh International World Wide Web Conference, Brisbane, Australia, 1997, ACM Press.
[8] L. Page, S. Brin, R. Motwani, T. Winograd, The pagerank citation ranking: bringing order to the web, Technical report, Stanford Digital Libraries, 1998.
[9] B. Amento, L. Terveen, W. Hill, in: Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Does ''Authority'' Mean Quality? Predicting Expert Quality Ratings of Web Documents, 2000.
[10] H. Borges, M. Cervi, P.T. Álvarez de Arcaya, G. Guardado, R. Rabaza, J. Sosa, Rate of compliance with the HON code of conduct versus number of inbound links as quality markers of pediatric web sites, in: Proceedings of the Sixth World Congress on the Internet in Medicine, Udine, Italy, 29 November—2 December 2001, http://mednet2001.drmm.uniud.it/proceedings/paper.php?id=75.
[11] C. Boyer, M. Selby, R.D. Appel, The health on the net code of conduct for medical and health web sites, 1997, MEDNET97—European Congress on the Internet in Medicine, Brighton, UK, 3—6 November 1997.
[12] M.A. Mayer, A. Leis, R. Sarrias, P. Ruiz, Web Mèdica Acreditada Guidelines: Reliability and Quality of Health Information on Spanish-language websites, vol. I, No. 1, in: R. Engelbrecht, et al. (Eds), Connecting Medical Informatics and Bioinformatics, Proceedings of the 19th International Congress of the European Federation for Medical Informatics, MIE 2005, Geneve, Switzerland, Munich-Heuherberg, Germany, 2005, pp. 1287—1292.
[13] G. D'Andrea, Health web site accreditation: opportunities and challenges, Manage. Care Q. 10 (1) (2002) 1—6.
[14] C. Boyer, A. Geissbuhler, A decade devoted to improving online health information quality, in: Proceedings of the 19th International Congress of the European Federation for Medical Informatics, MIE 2005, Geneve, Switzerland, Munich-Heuherberg, Germany, 2005.
[15] D. Fallis, M. Fricke, Indicators of accuracy of consumer health information on the Internet: a study of indicators relating to information for managing fever in children in the home, J. Am. Med. Inform. Assoc. 9 (1) (2002) 73—79.
[16] N. Miller, E.M. Lacroix, J.E. Backus, MEDLINEplus: building and maintaining the National Library of Medicine's Consumer Health Web Service, Bull. Med. Libr. Assoc. 88 (1) (2000) 11—17.
[17] G. Eysenbach, G. Yihune, K. Lampe, P. Cross, D. Brickley, A metadata vocabulary for self- and third-party labeling of health web-sites: health information disclosure, description and evaluation language (HIDDEL), Proc. AMIA Symp. (2001) 169—173.
[18] T.H. Haveliwala, Topic-sensitive PageRank: a context-sensitive ranking algorithm for web search, IEEE Trans. Knowl. Data Eng. 15 (4) (2003) 784—796.
[19] K.J. Sparck, Synonymy and Semantic Classification, Edinburgh University Press, 1986.
[20] A.R. Aronson, T.C. Rindflesch, A.C. Browne, Exploiting a large thesaurus for information retrieval, in: Proceedings of RIAO, 1994.
[21] A.R. Aronson, O. Bodenreider, H.F. Chang, S.M. Humphrey, J.G. Mork, S.J. Nelson, T.C. Rindflesch, W.J. Wilbur, The NLM indexing initiative, Proc. AMIA Symp. (2000) 17—21.
[22] A.R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap Program, Proc. AMIA Symp. (2001) 17—21.
[23] Computation of related articles: http://www.ncbi.nlm.nih.gov/entrez/query/static/computation.html.
[24] A. Gaudinat, C. Boyer, Automatic extraction of MeSH terms from MEDLINEs abstracts, Workshop on Natural Language Processing in Biomedical Applications, NLPBA, 2002, pp. 53—57.
[25] A. Névéol, V. Mary, A. Gaudinat, C. Boyer, A. Rogozan, S. Darmoni, A benchmark evaluation of the French MeSH indexers, in: 10th Conference on Artificial Intelligence in Medicine (AIME 05), 23—27 July 2005, Aberdeen, Scotland.
[26] B.L. Humphreys, D.A.B. Lindberg, Building the Unified Medical Language System, in: Proceedings of the 13rd SCAMC, IEEE Computer Society Press, 1989, pp. 475—480.
[27] A. Burgun, O. Bodenreider, Methods for exploring the semantics of the relationships between co-occurring UMLS concepts, Proc. Medinfo. (2001) 171—175.
[28] M. Joubert, S. Aymard, D. Fieschi, F. Volot, et al., ARIANE: Integration of Information Databases within a Hospital Intranet, Int. J. Med. Inf. 49 (1998) 297—309.
[29] A.T. McCray, The UMLS Semantic Network, in: Proceedings of the 13rd SCAMC, IEEE Computer Society Press, 1989, pp. 503—507.
[30] A. Gaudinat, M. Joubert, S. Aymard, L. Falco, C. Boyer, M. Fieschi, WRAPIN: new generation health search engine using UMLS knowledge sources for MeSH term extraction from health documentation, Medinfo (2004) 356—360.
[31] G. Salton, M. McGill, The SMART Retrieval System—Experiments in Automatic Document Retrieval, Prentice Hall, Englewood Cliff, 1971.
[32] Text REtrieval Conference (TREC): http://trec.nist.gov.
[33] P. Ruch, R. Baud, C. Chichester, A. Geissbühler, Latent argumentative structuring for extraction of key sentences, vol. I, No. 1, in: Proceedings of the 19th International Congress of the European Federation for Medical Informatics, MIE 2005, Geneve, Switzerland, Munich-Heuherberg, Germany, 2005.
[34] I. Tbahriti, C. Chichester, F. Lisacek, P. Ruch, Using argumentation to retrieve articles with similar citations: an inquiry into improving related articles search in the MEDLINE Digital Library, Int. J. Med. Inform., 2005, in press.
[35] P. Ruch, R. Baud, J. Marty, A. Geissbühler, I. Tbahriti, A.-L. Veuthey, Latent Argumentative Pruning for Compact MEDLINE Indexing, AIME (2005) 246—250.
[36] P. Domingos, M. Pazzani, On the optimality of the simple Bayesian classifier under zero-one loss, Machine Learn. 29 (2—3) (1997) 103—130.
[37] Y. Yang, J. Pedersen, A comparative study on feature selection in text categorization, in: Proceedings of the ICML, 14th International Conference on Machine Learning, 1997, pp. 114—121.
[38] L. McKnight, P. Srinivasan, Categorization of sentence types in medical abstracts, AMIA (2003).
[39] S. Teufel, M. Moens, Sentence extraction and rhetorical classification for flexible abstracts, AAAI Spring Symposium on Intelligent Text Summarization, 1998, pp. 89—97.
[40] M. Hearst, User interfaces and visualization, in: Modern Information Retrieval by R. Baeza-Yates and B. Ribeiro-Neto, Addison Wesley.1999.
[41] B.J. Jansen, A. Spink, How are we searching the World Wide Web? A comparison of nine search engine transaction logs, Inf. Process. Manage. 42 (2006) 248—263, in press.
[42] Vivisimo clustering engine, 2004, http://vivisimo.com, accessed Oct. 2005.

[43] R.B. Allen, P. Obry, M. Littman, An interface for navigating clustered document sets returned by queries, in: Proceedings of the ACM Conference on Organizational Computing Systems, 1993, pp. 166—171.
[44] D. Eichmann, M. Ruiz, P. Srinivasan, Cross-Language Information Retrieval with the UMLS Metathesaurus. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 1998.
[45] P. Ruch, R. Baud, A. Geissbühler, Evaluating and reducing the effect of data corruption when applying bag of words approaches to medical records, Int J Med Inform. 67 (1—3) (2002 Dec 4) 75—83.
[46] J. Crowell, Q. Zeng, L. Ngo, E.M. Lacroix, A frequency-based technique to improve the spelling suggestion rank in medical queries, J. Am. Med. Inform. Assoc. (2004).
[47] K. Atkinson, GNU ASpell, version 0.50.3, Produced by SourceForge.Net, available at: http://aspell.sourceforge.net/, accessed on: February 2004.
[48] A. Levenstein, Binary codes capable of correcting deletions, insertions and reversals, Soviet Phys., Doklandy 10 (1966).
[49] W.B. Cavnar, J.M. Trenkle, N-gram-based text categorization, in: Proceedings of SDAIR-94, Third Annual Symposium on Document Analysis and Information Retrieval.