U-MedSearch: A Meta Search Engine of Medical Content for Different Users and Learning Needs

Marco Alfano, Biagio Lenzitti, Giosuè Lo Bosco

Abstract: *More and more people use Internet to look for medical information for understanding and learning but different users, such as experts (e.g., physicians) and consumers (e.g., patients), have different needs and bring different levels of reading ability and prior knowledge. Generic and specific search engines and specialized health sites either do not exploit the whole web or overload users with information of different nature. On the contrary, it is important for a user to immediately find the information on the topic being explored that has the 'right' amount of information and level of complexity.*

This paper presents a meta search engine of medical information on the web, U-MedSearch, that, for any keyword(s) provides four different lists of terms, and in turn web pages, on the basis of the used language (consumer or expert) and correlation degree (strong or loose) with the keyword(s) thus facilitating the search and learning paths of the different types of learners.

Key words: *Health Search Engine, Biomedical Information Retrieval, Consumer Health Vocabulary, Medical Dictionary, e-Learning.*

INTRODUCTION

Nowadays, more and more people use Internet to seek health and medical information for understanding and learning [3], [4], [6]. Different users, such as a patient, a physician or a medical researcher have diverse needs when searching for health topics and bring different levels of reading ability and prior knowledge together with a different vocabulary [7], [10].

Generic search engines (like Google, Bing or Yahoo) work on the whole web but make generic searches often overloading the user with the provided amount of information. Moreover, they are not able to provide specific information to different types of users. On the other hand, specific search engines, like PubMed[25] or Quertle[26], only work on medical literature (mostly PubMed). They provide extracts from medical journals that are mainly useful for medical researchers and experts but do not consider all the information contained in the web that can often provide additional insights to the specific research domain being explored [5].

Another source of information comes from the specialized web sites oriented either to consumers (e.g., WebMD[27], Healthline[28] or MedlinePlus[29]) or professionals (e.g., Health on Net Foundation Select[30], Translating research into practice[31] or MDConsult[32]). Those sites contain very focused information but are mainly built by hand and then miss all the huge amount of information that is available on the web. Moreover, there is often a fee to be paid in order to use them.

Internet users looking for medical information on the web for educational purposes would greatly benefit from a search engine that provides them with the 'right' information they are looking for without getting 'lost' with the amount and quality of information that Internet provides [3], [4], [8]. To this end, we consider two different types of learners, i.e., non experts and experts and two types of search needs, i.e., search for basic information and search for specific information. We have developed a meta search engine of medical information on the web, U-MedSearch that, for any keyword(s), provides four different lists of terms, and in turn of web pages, on the basis of the used language (consumer or

[25] http://www.ncbi.nlm.nih.gov/pubmed/
[26] http://www.quertle.info/
[27] http://www.webmd.com/
[28] http://www.healthline.com/
[29] http://www.nlm.nih.gov/medlineplus/
[30] http://www.hon.ch/
[31] http://www.tripdatabase.com/
[32] http://www.mdconsult.com/

expert) and correlation degree (strong or loose) with the keyword(s) thus facilitating the search and learning paths of the different types of learners.

The paper is organized as follows. The second section describes the basic principles of the U-MedSearch methodology. The third section presents the architecture and implementation details of U-MedSearch together with some experimental results. The final section presents some conclusions and future work.

U-MEDSEARCH METHODOLOGY

As discussed above, users with different skills and learning needs use internet for finding the medical content of interest but the available tools are too generic (e.g., google) or too specific (e.g., PubMed or WebMD) and are not able to provide the user with the most suitable web information without overloading him/her. Thus a web search engine which provides the different users with the 'right' information and level of complexity would, undoubtedly, be of great benefit.

In what follows we consider two different types of learners, i.e., 'non-experts' (e.g., patients, school students) who do not use (and do not know) the medical-technical terminology and 'experts' (e.g., physicians, medical researchers) who use (and know well) the medical terminology. Moreover, we assume that users can search for either basic information, i.e., strongly correlated to the searched keyword(s), for example to understand the searched topic, or specific information, i.e., loosely correlated to the searched keyword(s), to deepen or expand their knowledge on the searched topic. This will lead to four different search categories and, as a consequence, learning paths:

1. Non-experts looking for basic information;
2. Non-experts looking for specific information;
3. Experts looking for basic information;
4. Experts looking for specific information.

Thus, starting from a search performed with one or more keywords, the goal is to find the main medical terms (made of one or more words) that are representative of each category so that the user can use such terms either directly, for his/her learning purposes, or to perform a more detailed search. To this end, we consider the *Unified Medical Language System® (UMLS)*[33]*,* the largest collection of multilingual vocabularies that contains information about biomedical and health related concepts, created and maintained by the 'US National Library of Medicine' and, in particular, the following vocabularies that are chosen because well cover the different medical terminologies:

- The *Medical Subject Headings (MeSH)*[34], developed and maintained by the 'U.S. National Library of Medicine', is used for indexing, cataloging, and searching for biomedical and health-related information and documents particularly into the world's leading biomedical journals for the MEDLINE/PubMED databases. MeSH is used by indexers, subject catalogers, online searchers and in retrieval systems.
- The *Medical Dictionary for Regulatory Activities (MedDRA)*[35], developed by the 'International Conference on Harmonisation' and owned by the 'International Federation of Pharmaceutical Manufacturers and Associations', contains medical terminology used for the specific use of sharing regulatory information and clinical safety data for human medical products. Users of MedDRA include pharmaceutical companies, biotechnology companies, device manufacturers, regulatory authorities, CROs, system developers, support service organizations, health care professionals, researchers and other interested parties outside of the regulated pharmaceutical/biological industry.

[33] http://www.nlm.nih.gov/research/umls/
[34] https://www.nlm.nih.gov/mesh/
[35] http://www.meddra.org/

- The *Open-access and collaborative (OAC) consumer health vocabulary (CHV)*[36], produced by the 'Biomedical Informatics Department at the University of Utah', connects informal, common words and phrases about health to technical terms used by healthcare professionals and found in the UMLS [9]. It presents, among others, two fields: the *CHV Preferred Name*, i.e., the preferred consumer term, and the *UMLS Preferred Name*, i.e., the preferred 'medical' term as defined by UMLS. This particular characteristic of the OAC-CHV has led us to consider it as two separated vocabularies, one containing the consumer terms, CHV Preferred Names, and we call it *CHV_P*, and the other one containing the technical counterparts, UMLS Preferred Names, and we call it *CHV_S*. Notice that CHV_P and CHV_S do not have terms in common.

Having these vocabularies at hand, we assume that a user starts his/her search on a topic, e.g., through a generic web search engine, by choosing an initial keyword(s). We then take the first n web pages returned by the search engine and extract all the medical terms present, at least, in one of the medical vocabularies presented above, i.e., MeSH, MedDRA, CHV_P and CHV_S.

Let X be the set of all extracted medical terms, we consider a partition of X, P(X), a set of nonempty subsets of X such that every element x in X only belongs to one of these subsets. In particular, considering that CHV_S and CHV_P are disjoint, we have the following subsets:

1. $\mathit{S\_MeSH} = \{x \in X \mid x \in \text{MeSH and } x \notin (\text{MedDRA} \cup \text{CHV\_P} \cup \text{CHV\_S})\}$
2. $\mathit{S\_MedDRA} = \{x \in X \mid x \in \text{MedDRA and } x \notin (\text{MeSH} \cup \text{CHV\_P} \cup \text{CHV\_S})\}$
3. $\mathit{S\_CHV\_P} = \{x \in X \mid x \in \text{CHV\_P and } x \notin (\text{MeSH} \cup \text{MedDRA})\}$
4. $\mathit{S\_CHV\_S} = \{x \in X \mid x \in \text{CHV\_S and } x \notin (\text{MeSH} \cup \text{MedDRA})\}$
5. $\mathit{S\_MeSH\_MedDRA} = \{x \in X \mid x \in (\text{MeSH} \cap \text{MedDRA}) \text{ and } x \notin (\text{CHV\_S} \cup \text{CHV\_P})\}$
6. $\mathit{S\_MeSH\_CHV\_P} = \{x \in X \mid x \in (\text{MeSH} \cap \text{CHV\_P}) \text{ and } x \notin (\text{MedDRA}) \}$
7. $\mathit{S\_MeSH\_CHV\_S} = \{x \in X \mid x \in (\text{MeSH} \cap \text{CHV\_S}) \text{ and } x \notin (\text{MedDRA}) \}$
8. $\mathit{S\_MedDRA\_CHV\_P} = \{ x \in X \mid x \in (\text{MedDRA} \cap \text{CHV\_P}) \text{ and } x \notin (\text{MeSH}) \}$
9. $\mathit{S\_MedDRA\_CHV\_S} = \{x \in X \mid x \in (\text{MedDRA} \cap \text{CHV\_S}) \text{ and } x \notin (\text{MeSH}) \}$
10. $\mathit{S\_MeSH\_MedDRA\_CHV\_P} = \{x \in X \mid x \in (\text{MeSH} \cap \text{MedDRA} \cup \text{CHV\_P}) \}$
11. $\mathit{S\_MeSH\_MedDRA\_CHV\_S} = \{x \in X \mid x \in (\text{MeSH} \cap \text{MedDRA} \cup \text{CHV\_S}) \}$

For each term, we determine which subset the term belongs to and count the occurrence number of the term in all pages. We then group the terms in the following way:

a. The terms found in the subset *S_MeSH_CHV_P ∪ S_MeSH_MedDRA_CHV_P* go to the "Non-experts looking for basic information" category;
b. The terms found in the subset *S_MedDRA_CHV_P* go to the "Non-experts looking for specific information" category;
c. The terms found in the subset *S_MeSH_MedDRA_CHV_S ∪ S_MeSH_CHV_S* go to the "Experts looking for basic information" category;
d. The terms found in the subset *S_MedDRA_CHV_S* go to the "Experts looking for specific information" category.

Fig. 1 graphically shows the eleven subsets and their relationship with the four categories.

[36] http://consumerhealthvocab.org/

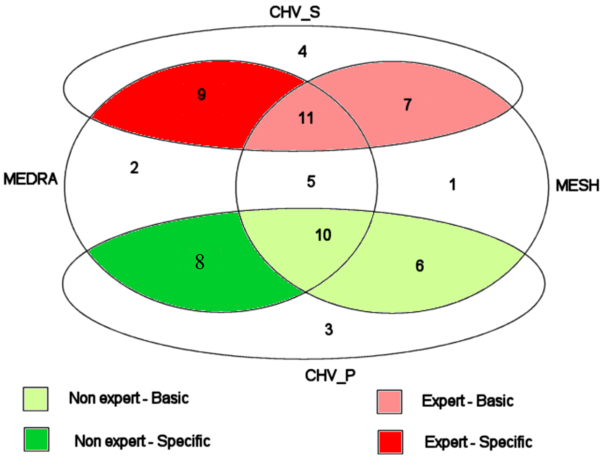


Figure 3. Venn diagram of the eleven subsets and four categories

For each term, we then create a specific correlation with the initial keyword(s) that will allow one type of navigation and, in turn, of learning path. Notice that we have chosen which subset to use for each category according to the following explanation:

- the terms that go into the "non-experts looking for basic information" category belong to the MeSH dictionary, indicating that they are part of a general medical terminology, belong to the CHV_P dictionary, indicating that are suited for non experts and are recurring (high number of occurrences) indicating that they have a strong correlation with the initial keyword(s). They will be used by a non-expert, for example, for understanding the searched topic (e.g., "insulin" and "blood sugar" for the "diabetes" keyword);
- the terms that go into the "non-experts looking for specific information" category belong to the MedDRA dictionary indicating that they are part of a specific medical terminology, belong to the CHV_P, indicating that are suited for non experts and are sporadic (low number of occurrences) indicating that they have a loose correlation with the initial keyword(s). They will be used by a non-expert, for example, to investigate in depth on the searched topic (e.g., "heart attack" and "pregnant" for the "diabetes" keyword);
- the terms that go into the "experts looking for basic information" category belong to the MeSH dictionary, indicating that they are part of a general medical terminology, belong to the CHV_S indicating that are suited for experts and are recurring (high number of occurrences) indicating that they have a strong correlation with the initial keyword(s). They will be used by an expert, for example, for understanding the searched topic (e.g., "diabetes mellitus" and "blood glucose" for the "diabetes" keyword);
- the terms that go into the "experts looking for specific information" category belong to the MedDRA dictionary, indicating that they are part of a specific medical terminology, belong to the CHV_S, indicating that are suited for experts and are sporadic (low number of occurrences), indicating that they have a loose correlation with the initial keyword(s). They will be used by an expert, for example, to investigate in depth on the searched topic (e.g., "hba1c" and "stress" for the "diabetes" keyword).

U-MEDSEARCH IMPLEMENTATION AND EXPERIMENTAL RESULTS

Starting from our previous work on the web search field [1], [2], we have implemented the methodology presented above. The system takes one or more keywords as input, retrieves the related web pages, and provides the four lists of terms belonging to the four categories introduced above. Fig. 2 shows the basic architecture of the system.

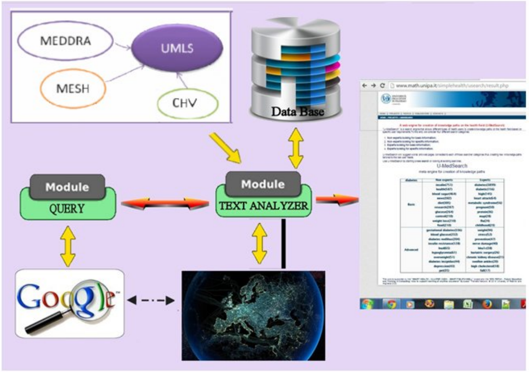
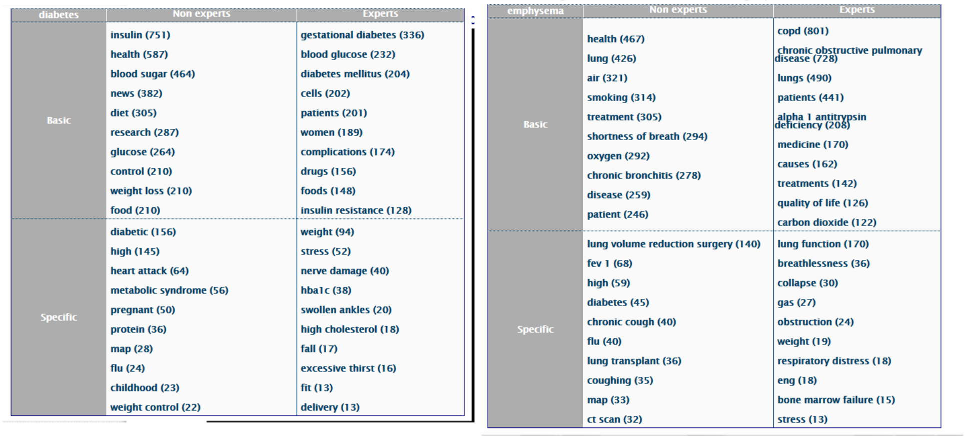


Figure 2. System architecture.

The QUERY module takes the keyword(s) and the n number of web pages to be analyzed. It then searches this keyword(s) through Google and takes the first n results creating a collection of n pages with related links. For each link, the TEXT ANALYZER module retrieves the related web page, cleans it, by removing tags and stop (common) words and then verifies whether each extracted term (word or combination of two or more consecutive words, e.g., 'diabetes mellitus') is contained in one of the medical vocabularies introduced before, i.e., MeSH, MedDRA, CHV_P and CHV_S. In this case, the analyzer stores the extracted term, the name of the partition subset which the term belongs to and the number of occurrences in the DB database, otherwise it disregards the term. At the end, we will have all the medical terms found in the web pages grouped by each subset they belong to and we then create the four term categories.

To test our system we ran some experiments assuming that users search for health terms (diseases, symptoms, treatments, etc.) in order to understand and learn more on them. We used various keywords and decided to analyze one hundred web pages for each keyword. For each search, our system provided us with four lists with the ten most recurring terms of each category. The results of the experiments for all the keywords can be found at the address http://www.math.unipa.it/simplehealth/umedsearch. Fig. 3 shows the terms of the four categories (together with the occurrence number) for "Diabetes" and "Emphysema" keywords.



diabetes	Non experts	Experts
Basic	insulin (751)	gestational diabetes (336)
	health (587)	blood glucose (232)
	blood sugar (464)	diabetes mellitus (204)
	news (382)	cells (202)
	diet (305)	patients (201)
	research (287)	women (189)
	glucose (264)	complications (174)
	control (210)	drugs (156)
	weight loss (210)	foods (148)
	food (210)	insulin resistance (128)
Specific	diabetic (156)	weight (94)
	high (145)	stress (52)
	heart attack (64)	nerve damage (40)
	metabolic syndrome (56)	hba1c (38)
	pregnant (50)	swollen ankles (20)
	protein (36)	high cholesterol (18)
	map (28)	fall (17)
	flu (24)	excessive thirst (16)
	childhood (23)	fit (13)
	weight control (22)	delivery (13)

emphysema	Non experts	Experts
Basic	health (467)	copd (801)
	lung (426)	chronic obstructive pulmonary disease (728)
	air (321)	lungs (490)
	smoking (314)	patients (441)
	treatment (305)	alpha 1 antitrypsin deficiency (208)
	shortness of breath (294)	medicine (170)
	oxygen (292)	causes (162)
	chronic bronchitis (278)	treatments (142)
	disease (259)	quality of life (126)
	patient (246)	carbon dioxide (122)
Specific	lung volume reduction surgery (140)	lung function (170)
	fev 1 (68)	breathlessness (36)
	high (59)	collapse (30)
	diabetes (45)	gas (27)
	chronic cough (40)	obstruction (24)
	flu (40)	weight (19)
	lung transplant (36)	respiratory distress (18)
	coughing (35)	eng (18)
	map (33)	bone marrow failure (15)
	ct scan (32)	stress (13)

Figure 3. Terms classification for "Diabetes" and "Emphysema" keywords

Notice that each term in the four lists is clickable and provides the web page with the highest number of occurrences of that term so allowing the user to immediately find a web page of interest without having to examine all google results one by one. Any user can then easily create a personalized learning path. For example, a medicine student, studying "emphysema", can use, as a first step, the "Chronic obstructive pulmonary disease

(COPD)" term and related web page for a basic learning and, as a second step, the "lung function" term and related web page for a deeper learning of the subject.

CONCLUSIONS AND FUTURE WORK

This paper has presented the methodology, implementation details and preliminary experimental results of U-MedSearch, that facilitates the search and learning paths of different types of learners by associating four different lists of terms to each keyword(s) on the basis of the used language (non expert or expert) and correlation degree (strong or loose). The experimental results are encouraging and show the effectiveness of our system in separating the 'non-expert' terms from the 'expert' ones and the 'basic' terms from the 'specific' one, thus allocating each term into the proper category. Nevertheless, a deeper understanding of the correlation between the terms and the initial keyword(s) needs further analysis (with experts and non experts). Moreover, the methodology needs to be refined (e.g., considering other vocabularies) and more experiments are necessary to increase the precision of the terms allocation to the four categories. Further experiments also need to be performed on the web pages so to visually verify they are really the best suggestions for each type of searcher.

ACKNOWLEDGMENT

This work has partially been funded by the PON Smart Cities PON04a2_C "SMART HEALTH – CLUSTER OSDH – SMART FSE-STAYWELL" project.

REFERENCES

[1] Alfano, M., Lenzitti, B., and Lo Bosco, G., "A web search methodology for health consumers", Proc. of CompSysTech'14, pp. 150–157, 2014.

[2] Alfano, M., and Lenzitti, B., "U-Search: A meta engine for creation of knowledge paths on the web", Proc. of CompSysTech'10, pp. 442–447, 2010.

[3] Cline, R. J., and Haynes, K. M., "Consumer health information seeking on the Internet: the state of the art", Health Educ. Res., vol. 16, no. 6, pp. 671–92, Dec. 2001.

[4] ECDC Technical Report, "A literature review on health information-seeking behaviour on the web: a health consumer and health professional perspective", 2011.

[5] Hersh, W., Information Retrieval: A Health and Biomedical Perspective, Springer, 2009.

[6] Pletneva, N., Vargas, A., and Boyer, C., "Requirements for the general public health search," Khresmoi Public Deliv., 2011.

[7] Seedorff, M., and Peterson, K., "Incorporating Expert Terminology and Disease Risk Factors into Consumer Health Vocabularies", Pac. Symp. Biocomp., pp. 421–432, 2013.

[8] Stvilia B., Mon L., and Yi, Y., "A model for online consumer health information quality," Journ. Am. Soc. for Inform. Sci. and Tech., vol. 60, pp. 1781–1791, 2009.

[9] Zeng, Q., and Tse, T. "Exploring and developing consumer health vocabularies", Journ. Am. Med. Inform. Ass., vol. 13, no.1, pp. 24–29, 2006.

[10] Zielstorff, R. D., "Controlled vocabularies for consumer health", Journ. Biomed. Inform., vol. 36, no. 4–5, pp. 326–333, Aug. 2003.

ABOUT THE AUTHORS

Marco Alfano, PhD, Anghelos Centre on Communication Studies and Dipartimento di Matematica e Informatica, University of Palermo, Palermo, Italy, Phone: +39 091 341791, E-mail: marco.alfano@anghelos.org.

Assist. Prof. Biagio Lenzitti, Dipartimento di Matematica ed Informatica, University of Palermo, Palermo, Italy, Phone: +39 091 23891101, E-mail: biagio.lenzitti@unipa.it.

Assist. Prof. Giosuè Lo Bosco, Dipartimento di Matematica ed Informatica, University of Palermo and I.E.ME.ST. Istituto Euro-mediterraneo di Scienza e Tecnologia, Palermo, Italy, Phone: +39 091 23891075, E-mail: giosue.lobosco@unipa.it.

The paper has been reviewed.