Design and Evaluation of the iMed Intelligent Medical Search Engine

Gang Luo
IBM T.J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532, USA
luog@us.ibm.com

Abstract — **Searching for medical information on the Web is popular and important. However, medical search has its own unique requirements that are poorly handled by existing medical Web search engines. This paper presents iMed, the first intelligent medical Web search engine that extensively uses medical knowledge and questionnaire to facilitate ordinary Internet users to search for medical information. iMed introduces and extends expert system technology into the search engine domain. It uses several key techniques to improve its usability and search result quality. First, since ordinary users often cannot clearly describe their situations due to lack of medical background, iMed uses a questionnaire-based query interface to guide searchers to provide the most important information about their situations. Second, iMed uses medical knowledge to automatically form multiple queries from a searcher' answers to the questions. Using these queries to perform search can significantly improve the quality of search results. Third, iMed structures all the search results into a multi-level hierarchy with explicitly marked medical meanings to facilitate searchers' viewing. Lastly, iMed suggests diversified, related medical phrases at each level of the search result hierarchy. These medical phrases are extracted from the MeSH ontology and can help searchers quickly digest search results and refine their inputs. We evaluated iMed under a wide range of medical scenarios. The results show that iMed is effective and efficient for medical search.**

I. INTRODUCTION

Today, both ordinary Internet users (6% of American Internet users on an average day) and doctors are increasingly using Web search engines (WSEs) to search for medical information on the Web [9], [31]. In response to this huge market need, several medical WSEs have been launched since October 2005. These systems provide useful information to searchers as background knowledge rather than making exact diagnosis [20]. They include Healthline [12], Google Health [10], SearchMedica [29], and Medstory [21]. Nevertheless, medical search has its own unique requirements that distinguish itself from traditional Web search. Existing medical WSEs have not fully addressed these unique requirements and cannot completely fulfill medical information searchers' needs.

All existing medical WSEs assume that searchers can form appropriate queries by themselves. However, most Internet users do not have much medical knowledge. Frequently, a medical information searcher has only a vague idea about the problem that he is facing and does not know the proper way to clearly describe his situation in sufficient detail. As a result, appropriate guidance is highly necessary during the medical search process. This can be illustrated by an analogy to the medical diagnosis process. During a doctor's office visit, the most important purpose of the conversation between the doctor and the patient is to let the doctor guide the patient to collect enough useful information about the patient's situation. The doctor asks a sequence of proper questions, where the content of the next question is based on the answers the patient gives to the previous questions. Then the doctor uses the collected information to perform differential diagnosis.

In deciding which diseases a patient is having, both the presence and the absence of certain symptoms can provide important clues (e.g., whether sputum is accompanied by coughing). Existing WSE technologies are mainly based on keyword matching. A searcher cannot easily describe the absence of a symptom S_y in a way that can be well utilized by existing WSEs. For example, Web pages describing the diseases that do not have S_y can either mention "without S_y" explicitly in many different ways or do not mention S_y at all. For a Web page P describing a disease D, the presence or absence of the keywords for S_y on P cannot be directly used as the criterion for deciding whether D has S_y. Hence, it is difficult for a medical information searcher to obtain useful search results purely through traditional keyword search.

To partially address the aforementioned problems, Healthline, a major medical WSE, added a symptom search functionality in Feb. 2007 [12], [24]. The searcher selects one or more symptoms from a given symptom list (no additional keywords are allowed), and then Healthline returns a list of diseases that have all these symptoms. Symptom search is helpful but only addresses the aforementioned problems in a limited way. First, many diseases have the same symptom while the correct way to distinguish these diseases is to use some features (e.g., sex, age, race, patient occupation) other than the remaining symptoms in the given symptom list [4]. It is extremely difficult and time-consuming for the searcher to distinguish these diseases himself by checking their detailed descriptions. For example, as mentioned in Collins [4], more than forty diseases can cause abdominal swelling. We can distinguish those diseases using only the location and some other properties of the swelling rather than any other symptoms. Second, symptom search cannot express the absence of certain symptoms, which can significantly reduce the length of the list of possible diseases. Third, symptom search disallows the searcher from inputting other useful keywords for identifying the possible diseases, such as exam results, existing diseases, and the foods, beverages, and medicines that the patient has taken. Fourth, a patient can have several symptoms simultaneously due to the presence of multiple diseases. Symptom search can only find those diseases that have all these symptoms, while those diseases can be completely different from the ones that the patient is having. Even worse, the returned disease set for multiple symptoms is often empty.

In this paper, we present iMed, a prototype *i*ntelligent *med*ical WSE that addresses the aforementioned limitations of existing systems. iMed introduces and extends expert system technology into the search engine domain. It is the first intelligent medical WSE that extensively uses medical knowledge and questionnaire to facilitate the search process. The idea of iMed is to mimic both the way that doctors interact with patients and doctors' differential diagnosis reasoning process, while maintaining existing medical WSEs' strength of handling keyword queries.

As mentioned before, the traditional keyword query interface of existing WSEs is unsuitable for medical search when the searcher does not know the proper way to clearly describe his situation in sufficient detail. Especially, the searcher often has no idea about what information is important for finding the desired results. In this case, we should guide the searcher to describe his situation appropriately so that iMed can collect enough useful information without receiving much junk information that interferes with the search process.

The key insight for the design of a guided medical search interface comes from our observation that doctors often use questionnaires to interact with patients [4], [28]. Questionnaires can easily guide the searcher to provide the most important information about his situation. They are user friendly and require no special user training, as they are frequently encountered in daily life. Nevertheless, a fixed questionnaire is too rigid. We should allow searchers to input other useful information that has not been addressed by the questions in the questionnaire. Based on these insights, we design our query interface as the combination of a questionnaire and traditional keyword text areas. The searcher answers questions in the questionnaire and inputs into text areas other useful information not covered by the questionnaire.

The searcher's answers to the questions should not be directly used as query keywords to perform medical search. In fact, straightforward keyword matching using these answers often leads to undesirable search results. This can be illustrated by an analogy to medical diagnosis. Medical diagnosis makes heavy use of medical knowledge and is a complex reasoning process. Existing keyword matching techniques cannot handle many important issues in this reasoning process, such as the absence of certain symptoms, some symptom properties (e.g., lasting time, degree), patient age, and quantitative test results. To obtain good medical search results, we need to use medical knowledge to transform these answers into appropriate keywords, and combine them with the other keywords that the searcher inputs into the text areas to form multiple queries. iMed uses all those queries, rather than a single query, to perform search.

If we treat the transformation from question answers into query keywords as an exact medical diagnosis process, then this transformation will seem like an almost impossible task according to our experience of repeated failures of medical expert systems in the past several decades [15], [36]. In practice, exact medical diagnosis is a complex process [15]. It requires much more information (e.g., detailed physical exam results and lab test results offered by medical professionals) than what an ordinary user can provide in a simple questionnaire. It also needs much practical experience and deep medical knowledge that cannot be easily mimicked using existing artificial intelligence techniques [36]. Fortunately, the purpose of using a medical WSE is to provide useful information to searchers as background knowledge rather than making exact diagnosis [20]. As long as the search results are relevant and helpful to searchers, the medical search process is considered as successful. Hence, during the transformation from question answers into query keywords, we only need to maximize the probability that the resulting query keywords can facilitate searchers to find useful information. Searchers can refine their inputs multiple times after reading the search results (e.g., by changing the keywords inputted into the text areas). This greatly reduces the difficulty of the transformation. In general, medical search is an iterative process while medical expert systems usually only give the user a single chance. Thus, we would expect that iMed can succeed more easily than medical expert systems.

The key insight for the aforementioned transformation comes from our analysis of the success of medical diagnosis reminder systems. In the last several years, such reminder systems have been adopted in routine use by doctors in many hospitals [15], [26] and are much more successful than medical expert systems. A reminder system provides a short list of all possible diseases and useful tests based on a doctor's simple inputs so that he can quickly check which diseases and tests he has forgotten to consider (not much guidance is provided during the information input process). The success of these reminder systems largely relies on the fact that the first several steps of doctors' differential diagnosis reasoning process are usually based on some empirical rules [4], [13], [28]. In other words, it is often feasible to use simple pattern matching to greatly reduce the number of diseases and tests that need to be considered without making exact diagnosis. Based on this insight, we use decision trees written by medical professionals to facilitate our transformation.

In practice, a medical information searcher is often uncertain about his exact medical problems and prefers to learn all kinds of knowledge related to his situation [5], [20], [25]. For this purpose, iMed forms multiple queries to perform search simultaneously and structures all their search results into a multi-level hierarchy that has explicitly marked medical meanings [19]. Using this hierarchy, searchers can efficiently navigate among the search results and quickly obtain desired information.

One way to structure the search results into a hierarchy is to use automatic clustering methods. However, these methods are unsuitable for intelligent medical search, because they were designed for the open domain without utilizing any specific knowledge. In the medical domain, we have domain-specific knowledge and searchers' desired (sub-)categories are generally known in advance. The (sub-)categories obtained by automatic clustering methods usually do not match with the (sub-)categories desired by medical information searchers.

To ensure that the (sub-)categories in the search result hierarchy match with medical information searchers' expectations, we use a novel automatic query formation method. The overview Web page for each category is retrieved using a query specifically formed for the corresponding topic. Also, the search results in each sub-category of a category are obtained using a query specifically formed for the corresponding aspect and topic. When forming these queries, we use both searchers' inputs and medical knowledge, while considering the different roles that various levels play in the search result hierarchy.

Good medical WSEs should automatically suggest related medical phrases to help searchers quickly digest search results and refine their inputs [12], [20], [38]. At each level of the search result hierarchy, iMed generates a single candidate set of related medical phrases for all the formed queries by considering their different weights. In ranking those medical phrases, iMed matches their representative Web pages [20] with the top Web pages retrieved for the queries.

In the traditional information retrieval literature, searchers input queries and the focus is on retrieving search results using better retrieval models. In contrast, this work focuses on automatically forming proper queries to obtain desired search results and effectively organizing these search results.

We crawled a large number of medical Web pages from the Internet and evaluated the effectiveness of our techniques under a wide range of medical scenarios. Our results show that iMed can perform medical search efficiently. Our experiments also show that user satisfaction is crucially tied to iMed's capability of guiding searchers to provide the most important information about their situations, automatically forming queries, constructing the search result hierarchy, and suggesting diversified, related medical phrases.

Besides medical search, the general ideas of this paper could also be applicable to other domain-specific (e.g., product) search. Suppose we have a knowledge base for a particular domain. Based on the criteria (e.g., price) specified by the searcher, this knowledge base can provide all those entities (e.g., cameras) satisfying these criteria and their interesting aspects (e.g., performance, design). Then an intelligent WSE for that domain can use our techniques to automatically form multiple queries and to build a multi-level search result hierarchy with explicitly marked meanings specific to that domain.

iMed is a sophisticated, evolving system with multiple technical components. In the past, we have reported its hierarchical search result output interface [19] and its iterative search advisor for facilitating iterative search [17], [18]. For readability and completeness, [17], [18], [19] have briefly presented a high-level overview of iMed. In this paper, we introduce several new technical components: questionnaire-based query interface, automatic query formation, and related medical phrase suggestion.

The rest of the paper is organized as follows. Section II describes iMed's user interface. Section III presents our search techniques. Section IV evaluates the effectiveness of our method under a wide variety of medical query scenarios. We conclude in Section V.

II. USER INTERFACE

The user interface of iMed contains two parts: the query interface and the answer interface. Both parts of iMed are different from those of existing medical WSEs.

A. Query Interface

In this section, we describe the query interface. In practice, we would expect most users of iMed to be ordinary Internet users without much medical knowledge, while medical professionals can also use iMed to help them accomplish their tasks [9], [13]. In designing iMed's query interface, we adopt the following principles to provide the greatest convenience to medical information searchers:

Principle 1: Minimize searchers' efforts.

Principle 2: Be easily accessible to ordinary users without much medical knowledge.

Principle 3: Be tolerant of imprecise user inputs.

Principle 4: Allow incomplete inputs.

We will illustrate these four principles when we describe iMed's query interface in detail.

1) *Query Interface Overview:* Fig. 1 shows the first screen of iMed's query interface. There are two possible cases:

Case 1: If the medical information searcher knows the appropriate query keywords (e.g., the exact name of the disease, the medicine, the test, the procedure, or the treatment), he can use the traditional keyword search interface to find desirable search results. In this case, iMed works in the same way as existing medical WSEs.

Case 2: If the searcher does not know the appropriate query keywords, he can use the questionnaire-based interface that is unique to iMed to guide him through the search process. In this case, the techniques used in iMed complement the techniques used in existing medical WSEs, as iMed uses medical knowledge to form keyword queries to perform search (see Section III).

Here, we focus on the second case.

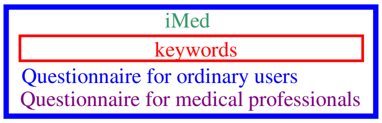

iMed
keywords
Questionnaire for ordinary users
Questionnaire for medical professionals

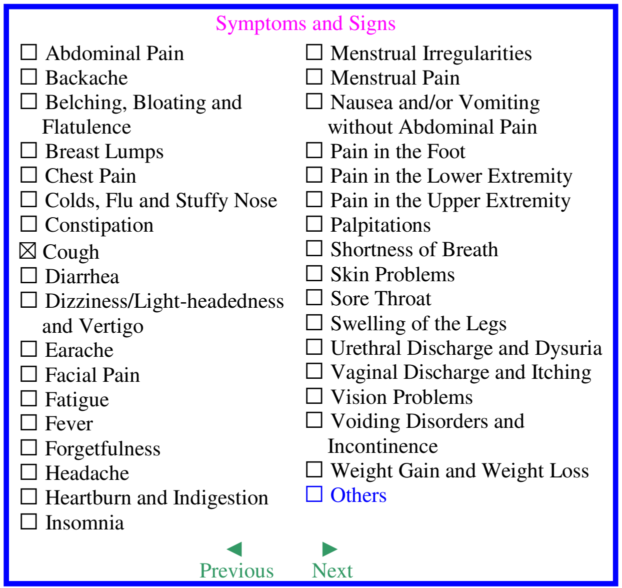
Fig. 1. The first screen of the query interface.

Since ordinary users and medical professionals have different capabilities of understanding medical terminology, iMed provides them with distinct questionnaire-based interfaces suitable for their various backgrounds [21] (Principle 2). The questionnaire for ordinary users mainly asks for symptoms and is written in daily language. Its content is based on the medical textbook written by Collins [4]. As shown in Sections II-A-2 and II-A-3, all the questions in it can be easily understood by ordinary users and do not require special exam and test results that can only be offered by medical professionals.

The questionnaire for medical professionals asks for exam and test results extensively and is largely written in medical terminology. Its content is based on the medical textbook written by Healey and Jacobson [13]. As mentioned in that book, such a questionnaire is welcome by medical professionals. Compared to the questionnaire for ordinary users, this questionnaire is more accurate for diagnosis purpose and can handle more difficult cases, while it often needs to ask more questions to significantly narrow down the list of possible diseases.

In the rest of the paper, we focus on the questionnaire for ordinary users. The questionnaire for medical professionals is similar and omitted. In the questionnaire, the searcher first selects subjective symptoms (e.g., fatigue) and objective signs (e.g., hypertension), and then answers questions about their detailed descriptions. The searcher can also input other useful information that is not covered by the questions into text areas.

2) *Symptoms and Signs:* iMed's questionnaire for ordinary users currently covers all the 267 symptoms and signs described in Collins [4]. It would be overwhelming to display all these symptoms and signs to searchers on a single page. Instead, iMed organizes this questionnaire into two levels. As shown in Fig. 2, the first level of this questionnaire contains the 34 most frequently encountered symptoms and signs accounting for more than 80% of the chief complaints with which physicians are confronted [28], and an "others" option. All the other 233 symptoms and signs described in Collins [4] are included in the "others" option as the second level of this questionnaire. To facilitate search, iMed classify those 233 symptoms and signs into multiple categories based on the affected body parts (e.g., general, head, neck, chest, abdomen, back, pelvic, extremities, skin). In most cases, the searcher can quickly find the appropriate symptoms and signs by checking only the first level of the questionnaire (Principle 1).



Symptoms and Signs

□ Abdominal Pain
□ Backache
□ Belching, Bloating and Flatulence
□ Breast Lumps
□ Chest Pain
□ Colds, Flu and Stuffy Nose
□ Constipation
☒ Cough
□ Diarrhea
□ Dizziness/Light-headedness and Vertigo
□ Earache
□ Facial Pain
□ Fatigue
□ Fever
□ Forgetfulness
□ Headache
□ Heartburn and Indigestion
□ Insomnia
□ Menstrual Irregularities
□ Menstrual Pain
□ Nausea and/or Vomiting without Abdominal Pain
□ Pain in the Foot
□ Pain in the Lower Extremity
□ Pain in the Upper Extremity
□ Palpitations
□ Shortness of Breath
□ Skin Problems
□ Sore Throat
□ Swelling of the Legs
□ Urethral Discharge and Dysuria
□ Vaginal Discharge and Itching
□ Vision Problems
□ Voiding Disorders and Incontinence
□ Weight Gain and Weight Loss
□ Others

Fig. 2. The first level of the questionnaire.

For each of the 267 symptoms and signs covered in the questionnaire, if its name is written in Collins [4] in medical phrases unfamiliar to ordinary users, we use the consumer health vocabulary [39] to annotate its name with layman terms. For example, the symptom "hemoptysis" is explained as "coughing up blood." As described in Zeng and Tse [39], the consumer health vocabulary is constructed from medical WSE query logs. It provides a mapping between medical phrases and layman terms frequently used by medical information searchers. Ordinary users can easily understand all the symptoms and signs written in layman terms in the questionnaire (Principle 2).

From all the 267 symptoms and signs in the questionnaire, the searcher can choose multiple of them reflecting his situation. Generally, when a doctor conducts medical diagnosis, he first identifies the chief complaints among all the patient's symptoms and signs (often there is only one chief complaint) and then performs analysis mainly based on these chief complaints [15]. However, in medical search, ordinary users usually have no rigorous medical training and cannot correctly identify their chief complaints [15]. To address this issue and to avoid missing important search results, iMed allows searchers to select multiple symptoms and signs without specifying their chief complaints (Principles 1 and 3).

3) *Question Pages:* For each of the 267 symptoms and signs covered in the questionnaire, Collins [4] has a companion diagnostic decision tree T_d. Each leaf node N of T_d contains the disease names that are most relevant to the branching conditions (in the non-leaf, non-root nodes) leading to N. iMed uses these diagnostic decision trees to prepare questions for the symptoms and signs and also to transform question answers into query keywords. In this section, we show how questions are generated one by one using these trees. In Section III, we show how to transform question answers into query keywords.

After obtaining all the symptoms and signs chosen by the searcher, iMed will generate question pages to ask questions about their detailed descriptions. Each question page contains one or more questions. The questions in the next question page are selected according to the answers the searcher provides to the questions in the previous question pages, as if iMed were traversing the corresponding diagnostic decision trees for these symptoms and signs. iMed can display all the used diagnostic decision trees on the answer interface and highlight the traversed paths to facilitate the searcher in understanding the underlying medical reasoning process.

For example, Fig. 3 shows the diagnostic decision tree for the symptom cough that is described in Collins [4]. If cough is the only symptom chosen by the searcher, the first question page generated by iMed will contain a single question "Is there significant sputum production?", as shown in Fig. 4. If the searcher answers "yes" to this question, iMed's next question will be "Is the sputum purulent?" Otherwise if the searcher answers "no" to this question, iMed's next question will be "Do you have difficulty breathing?"

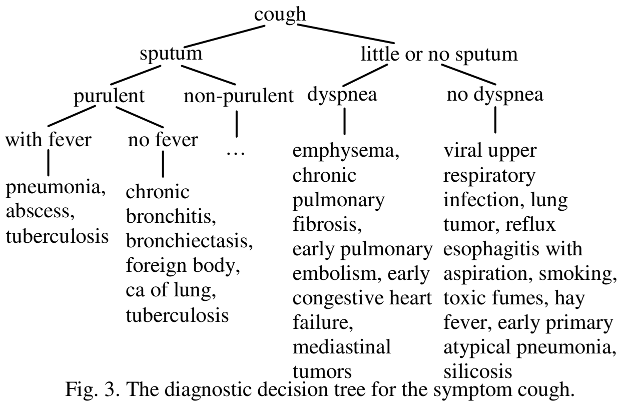
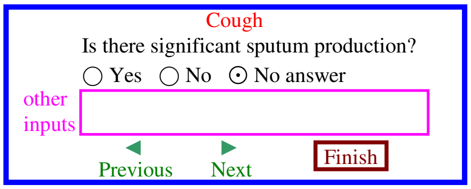

cough
sputum
little or no sputum
purulent
non-purulent
dyspnea
no dyspnea
with fever
no fever
…
pneumonia, abscess, tuberculosis
chronic bronchitis, bronchiectasis, foreign body, ca of lung, tuberculosis
emphysema, chronic pulmonary fibrosis, early pulmonary embolism, early congestive heart failure, mediastinal tumors
viral upper respiratory infection, lung tumor, reflux esophagitis with aspiration, smoking, toxic fumes, hay fever, early primary atypical pneumonia, silicosis

Fig. 3. The diagnostic decision tree for the symptom cough.



Cough

Is there significant sputum production?

○ Yes ○ No ⊙ No answer

Fig. 4. The first question page that is generated for the symptom cough.

In generating questions, iMed uses the consumer health vocabulary [39] to rewrite difficult medical phrases in diagnostic decision trees into layman terms (Principle 2). For example, “dyspnea” in Fig. 3 is rewritten into “difficulty breathing.” Also, iMed may ask qualitative measures in the format of quantitative numbers and then convert these numbers into qualitative measures in order to traverse diagnostic decision trees. For instance, “Do you have fever?” can be asked as “What is your body temperature?”

For each question asked by iMed, the searcher can either answer it or provide no answer, as iMed allows incomplete inputs (Principle 4). In the case that the searcher provides no answer to a question Q_u, iMed may use some “backup” question to replace Q_u, as the diagnostic decision tree for a symptom or sign is generally not unique [1], [11], [14]. For other useful information that is not covered by the questions, the searcher can input its keywords into the “other inputs” text area that appears on every question page. The searcher can stop answering questions and obtain search results at any time by clicking the “finish” button that appears on every question page (Principle 4). In general, the more questions a searcher answers, the more information iMed has about his situation and the better the search results will be.

A question page can contain more than one question in the following two cases. First, if the searcher chooses multiple symptoms and signs, iMed will ask questions about all of them. Second, some nodes in certain diagnostic decision trees have multiple descendant branches with non-conflicting conditions. When iMed reaches a node N in a tree, if the searcher either provides no answer to the corresponding question or selects multiple answers simultaneously, iMed cannot traverse along a single descendant branch of N and has to ask corresponding questions for all the (selected) descendant branches of N.

When generating questions, iMed checks for redundancy to ensure that each same question is asked at most once. For example, at the first level of the questionnaire, if the searcher selects both symptoms cough and fever, fever will not be asked again when iMed generates questions for the symptom cough (see Fig. 3). Also, iMed only asks “consistent” questions. For instance, suppose the searcher selects a single symptom cough at the first level of the questionnaire. If he provides no answer to the question “Is there significant sputum production?”, iMed will not ask questions about sputum properties, such as “Is the sputum purulent?” Instead, iMed treats all such questions as if the searcher provided no answer. All the redundancy and consistency checking in the question generation process is coded as rules.

Most diagnostic decision trees written in Collins [4] have depths smaller than five. Thus, iMed will usually stop asking questions and produce search results in fewer than five question pages. This fulfills Principle 1 of minimizing searchers’ efforts.

B. Answer Interface

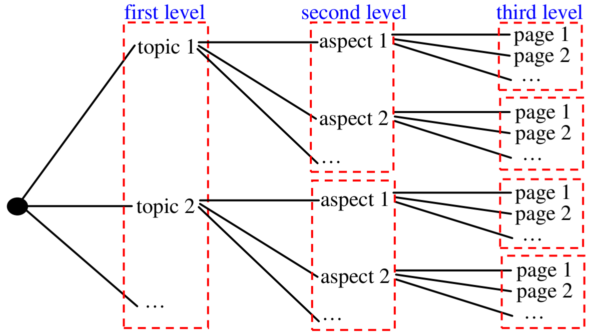


Fig. 5. The hierarchical structure of the answer interface.

For completeness, this section briefly summarizes iMed’s answer interface, which is presented with more details in [19]. In general, searchers prefer to simultaneously see various topics (e.g., disease names) that are potentially relevant to their medical situations [19]. For each such topic, searchers prefer to simultaneously see all kinds of aspects (e.g., symptom, diagnosis, and treatment) of it. As mentioned in Section III, iMed uses diagnostic decision trees to find those topics that are potentially relevant to the searcher’s medical situation. After obtaining the search results on those topics, iMed structures these search results into a three-level hierarchy that has explicitly marked medical meanings [19] to fulfill the above requirements. This is shown in Fig. 5.

At the first level of the hierarchy, all the search results are organized into multiple categories according to their topics (e.g., disease names). For each such topic T, an overview Web page P_T^o is provided to help the searcher determine whether this topic is related to his medical situation. At the second level, within each category, the corresponding search results are further divided into multiple sub-categories according to their aspects (e.g., symptom, diagnosis, and treatment). For each such aspect A of a topic T, an overview Web page $P_{A,T}^o$ is provided to help the searcher determine whether this aspect is related to his medical situation, while the retrieved Web pages are listed at the third level. At present, the topics mentioned at the first level of the hierarchy cover only diseases, while these topics can be easily expanded to include other contents (e.g., exams). To help a searcher quickly digest search results and refine his inputs (e.g., the keywords in the "other inputs" text areas) [12], [20], [38], at each level of the search result hierarchy iMed suggests to him a few medical phrases related to his medical situation.

III. SEARCH TECHNIQUES

iMed is a vertical WSE that crawls Web pages from a few selected, high-quality medical Web sites rather than all the Web sites. This approach is also adopted by MedSearch [20], Healthline [12], and Google Health [10] to avoid the disturbance of low-quality pages from irrelevant Web sites in the search results. Like MedSearch [20], iMed uses the Medical Subject Headings (MeSH) ontology [22] to identify medical phrases in the returned top Web pages and to rank medical phrases based on their relevance to searchers' inputs. MeSH is a standard vocabulary edited by the National Library of Medicine and widely used for indexing and cataloging biomedical and health-related documents.

Let C denote the collection of all the Web pages crawled by iMed. As standard pre-processing steps in Web information retrieval [20], for the Web pages in C, (1) all the HTML comments, JavaScript code, tags, and non-alphabetic characters are removed, (2) stopwords are dropped using the standard SMART stopword list [32], (3) noisy information is deleted using the frequent term sequence method described below, and (4) a forward index I_f and an inverted index I_i are built using the single-term vocabulary (i.e., the set of all the distinct words). In addition, another forward index I_f' that contains only medical phrases is built for the Web pages in C. iMed uses I_f' to suggest related medical phrases.

In a given Web site, the useful information in the Web pages is often accompanied by a lot of noisy information, e.g., navigation panels, copyright notices, and advertisements. Removing this noisy information can greatly improve both the quality of search results and the search speed [37]. We notice that a piece of noisy information usually appears in many Web pages and use the following frequent term sequence method to drop noisy information. A frequent term sequence is defined as a continuous sequence of terms that appears in many Web pages. For each Web page in the Web site, we obtain all the term sequences that contain n_1 continuous terms. All such term sequences that appear in more than n_2 Web pages are treated as frequent term sequences that represent noisy information. For each Web page P in the Web site, we identify all the frequent term sequences and remove them from P. Some of these frequent term sequences can have overlapping terms. In our current implementation of iMed, the default values of n_1 and n_2 are *6* and *15*, respectively.

After obtaining the searcher's answers to the questions, iMed proceeds in the following steps:
Step 1: Find the potentially relevant topics.
Step 2: Construct the search result hierarchy.
Step 3: Suggest related medical phrases.

A. Step 1: Finding Topics

In the questionnaire-based query interface of iMed, searchers do not input queries. Instead, iMed forms queries automatically based on searchers' inputs. More specifically, iMed uses medical knowledge to transform the searcher's question answers into several potentially relevant topics (diseases). For each such topic, iMed forms multiple queries to construct the corresponding part of the search result hierarchy.

We first show how to find the potentially relevant topics. The searcher chooses one or more symptoms and signs in the questionnaire, and iMed selects their diagnostic decision trees. For each such symptom or sign, iMed traverses to one or more branches in the corresponding diagnostic decision tree T_d based on the searcher's answers to the questions. Each leaf node of T_d contains several disease names [4]. The disease names in the leaf nodes of all these branches form a first set S_1 of medical phrases, and the disease names in all the other leaf nodes of T_d form a second set S_2 of medical phrases. A medical phrase M can appear in both S_1 and S_2 if the corresponding disease name appears in multiple leaf nodes of T_d. In this case, M is dropped from S_2.

Consider a selected diagnostic decision tree T_d. In general, all the branching conditions (e.g., symptoms, disease histories) in T_d have false positives and false negatives in diagnosing diseases [4], [13]. Moreover, searchers without much medical background can answer questions incorrectly due to unawareness of the exact medical definitions of these branching conditions. According to the medical diagnosis principles described in Collins *et al.* [4], [13], the medical phrases in both sets S_1 and S_2 can be relevant to the searcher's situation. The medical phrases in S_1 are generally more relevant than the medical phrases in S_2. Also, diseases not in T_d are usually irrelevant to the searcher's situation. To reflect the relevance of different medical phrases in T_d, we assign a local weight 1 to each medical phrase in S_1, and a local weight w_l<1 to each medical phrase in S_2. All the medical phrases in S_2 have the same local weight, as any branching condition can cause errors. The default value of w_l in iMed is 0.4.

Now consider all the selected diagnostic decision trees. A patient can have multiple symptoms and signs concurrently due to the presence of one or more diseases. To avoid omitting possible diseases, iMed needs to consider the set E of medical phrases from all these trees. As a general differential diagnosis principle [15], a medical phrase is more relevant to the searcher's situation if it is related to multiple symptoms and signs chosen by the searcher and appears in their diagnostic decision trees simultaneously. To consider this factor, for each medical phrase $M \in E$, iMed computes M's global weight as the sum of M's local weights in all the selected trees. This global weight reflects M's relevance to the searcher's situation. All the medical phrases in E are sorted in descending order of their global weights. In this way, the searcher's question answers are transformed into appropriately sorted medical phrases, and the searcher can find multiple relevant diseases (possibly for different symptoms and signs) simultaneously. (For all the diseases with the same global weight, we can sort them by their incidence rates [6], [7], [15] if such data are available. For a given disease, its incidence rate is the number of new cases per 1,000 people per year and reflects the probability of developing it.) Each such medical phrase is a potentially relevant topic.

B. Step 2: Constructing the Result Hierarchy

In this section, we discuss how to construct the search result hierarchy in iMed's answer interface. One might consider using classification to do this. For example, all the Web pages retrieved for a topic can be classified according to their aspects. However, online classification of search results is time-consuming and generally unsuitable for an interactive medical WSE like iMed. Also, it is difficult to know how many Web pages need to be retrieved for a topic *T* in order to obtain a sufficient number of search result Web pages for each aspect of *T*. Actually, even if we use a query formed for *T* to retrieve a large number of Web pages, it is still possible that no Web page among them mentions certain aspects of *T*.

1) *Overview:* To address the above problem, we use a novel automatic query formation method to construct the search result hierarchy. Our main observation is that the medical domain is a closed one. In the desired search result hierarchy, we can know the keywords for all the topics and their corresponding aspects. As a result, for each part of the search result hierarchy, we can use a different, specifically formed query to obtain the corresponding search result Web pages. More specifically, for each topic *T*, the overview Web page is retrieved using a query specifically formed for *T*. Also, for each aspect *A* of *T*, the corresponding search result Web pages are obtained using a query specifically formed for *A* of *T*. When forming these queries automatically, we use medical knowledge and consider the different roles that various levels play in the search result hierarchy. This can expedite the speed that searchers find their desired information. The resulting search result hierarchy fulfills all the requirements mentioned in Section II-B.

In our query formation method, we could form the complete set of queries for all found topics and all their aspects, use these queries to retrieve all the search results, and construct the entire search result hierarchy in a single batch. Nevertheless, this approach puts unnecessary burden on iMed and is undesirable. Searchers often skip completely many topics and aspects that they think are irrelevant to their medical situations at their first glance. Hence, there is no need to generate the search results for those topics and aspects. Moreover, searchers prefer to see iMed's outputs as soon as possible instead of waiting until the entire search result hierarchy has been constructed.

To reduce the load on iMed and to maximize the speed that searchers can see iMed's outputs, iMed constructs the search result hierarchy one part at a time. Each part of the hierarchy is generated only at the time that it is needed. If a part is never needed, it is never generated. More specifically, at the beginning iMed constructs only the first level of the search result hierarchy. If the searcher clicks a button and asks for more information about topic *T*, then iMed constructs for *T* the corresponding part of the second level of the search result hierarchy. Similarly, if the searcher clicks a button and asks for more information about aspect *A* of *T* at the second level, then iMed constructs for *A* of *T* the corresponding part of the third level of the search result hierarchy.

When constructing the search result hierarchy, we frequently encounter the case that multiple formed queries share a few common terms. In this case, we share the inverted list union computation task that is common to processing these queries. Consequently, processing these queries together is much faster than processing these queries separately. Below we describe the automatic query formation method in detail.

2) *First Level of the Hierarchy:* In this section, we discuss how to construct the first level of the search result hierarchy. As mentioned in Section III-A, the topics mentioned at the first level of the search result hierarchy cover the diseases in the leaf nodes of the selected diagnostic decision trees. All these diseases are sorted in descending order of their weights. For each such disease (topic) *T*, iMed provides an overview Web page P_T^o to help the searcher determine whether *T* is related to his medical situation.

We first consider the case that the searcher inputs no keyword into the "other inputs" text area (see Section II-A-3). To obtain the overview Web page P_T^o for disease *T*, we could use *T* as a query *Q* and treat the retrieved first result Web page P_1 as P_T^o. Nevertheless, this method is unsatisfactory because *Q* excludes many useful keywords related to the searcher's medical situation (i.e., selected symptoms and signs, answers to iMed's questions). As a result, P_1 may not contain those keywords and hence the searcher is unlikely to think that the description in P_1 (and thus *T*) matches with his medical situation, even if *T* is indeed his disease.

To address this problem, we combine all the available information about the searcher's medical situation into a query Q_T. Then we use the first result Web page retrieved by Q_T as the overview Web page P_T^o for disease *T*. The snippet of P_T^o is obtained using both Q_T and standard passage retrieval techniques [16].

More specifically, for a disease *T*, we form a query Q_T that includes the following three sets of information:

Set 1 (disease set): *T*.

Set 2 (symptom set): The symptoms or signs that are selected by the searcher and whose diagnostic decision trees contain *T*.

Set 3 (answer set): Some of the searcher's answers to iMed's questions. Each such answer A_n satisfies two conditions simultaneously. First, *T* is in a leaf node that is a descendant of A_n in the corresponding diagnostic decision tree. This ensures that A_n is relevant to *T*. Second, A_n describes either the presence of some symptom (e.g., "hypertension") or some fact that does not rely on numerical values (e.g., "constant pain"). This ensures that existing keyword matching techniques can well utilize A_n's keywords. We disregard any answer that describes either the absence of certain symptom (e.g., "no hypertension") or some fact relying on numerical values (e.g., "pain lasts two minutes or less"). For example, a Web page describing a disease that does not have symptom *S* can either mention the absence of *S* in many different ways or do not mention *S* at all. Consequently, existing keyword matching techniques cannot retrieve appropriate Web pages for a query that includes an answer describing the absence of *S*.

We use the example in Section II-A-3 to help readers understand the above query formation procedure. Suppose "cough" is the only symptom chosen by the searcher. The searcher answers "no" to iMed's first question "Is there significant sputum production?", and "yes" to iMed's second question "Do you have difficulty breathing (dyspnea)?". Then the keywords of the query formed for the disease "emphysema" are "emphysema cough difficulty breathing," while the keywords of the query formed for the disease "pneumonia" are "pneumonia cough."

In general, it is undesirable to form query Q_T as a simple combination of the keywords of the above mentioned three sets. Otherwise the answer set and the symptom set can contribute too many keywords that will overwhelm the keywords of the disease set. As a result, the retrieved overview Web page P_T^o is only marginally relevant to disease *T* and cannot serve well the purpose of helping the searcher determine whether *T* is related to his medical situation.

To address this problem, we use a weight-constrained method to limit the contributions from the symptom set and the answer set in the formed query Q_T. A few medical information searchers pointed out that the information in the disease set is generally more important than the information in the symptom set, which is generally more important than the information in the answer set. To reflect this point, we give each of the three sets a different weight: w_1 for the disease set, w_2 for the symptom set, and w_3 for the answer set, while $w_1 > w_2 > w_3$. For each keyword of a set, we give it a weight as follows. Consider a set *s* with weight *w*. After stopword removal, *s* has *n* keywords. Then each keyword of *s* obtains an equal proportion of *w*: *w/n*. In forming Q_T, we use all these three sets' keywords and consider their weights in the following way. For each distinct keyword *t* that appears in Q_T and whose total weight in these three sets is w_t, iMed treats *t* as if it appeared in Q_T w_t times. In iMed, the default weight values are: $w_1 = 1$, $w_2 = 0.7$, and $w_3 = 0.5$. A searcher can adjust those weight values according to his preference and inputs.

Next, we consider the case that the searcher inputs one or more keywords into the "other inputs" text area. In this case, we have a fourth set of information about the searcher's medical situation: the **other-inputs set**. A few medical information searchers pointed out that the information in this set is of the same importance as the information in the answer set. Thus, we give this set the same weight as the answer set: w_3. When constructing the first level of the search result hierarchy, we consider that weight and include the keywords of the other-inputs set in each formed query.

At the first level of the search result hierarchy, iMed's answer interface organizes all found topics into one or more *topic pages*, each of which contains ten *topic elements* for ten different topics. A button entitled "Without other inputs" is added on the bottom right corner of each topic page. If the searcher clicks this button, iMed returns to him a search result hierarchy as if no keyword were inputted into the "other inputs" text area.

Searchers can input arbitrary keywords into the "other inputs" text area. These keywords can provide hint to some diseases that searchers have but are not covered in the diagnostic decision trees corresponding to searchers' selected symptoms and signs. For example, if a searcher selects the symptom "chest pain" and inputs "cocaine" into the "other inputs" text area, then the query "chest pain cocaine" can retrieve a Web page about the disease "cocaine-induced myocardial ischemia" that is not covered in the diagnostic decision tree for "chest pain." To avoid missing such diseases, we include at the top of the first topic page an extra topic element *E*. Both the overview Web page and other search results of *E* are retrieved using a query Q_T that contains the keywords of both the symptom set and the other-inputs set.

3) *Second Level of the Hierarchy:* In this section, we discuss how to construct the second level of the search result hierarchy. When the searcher reaches the *aspect page* P_a for disease *T* at the second level of the search result hierarchy, he usually has read the overview Web page P_T^o for *T* at the first level of the search result hierarchy. There are two possible scenarios:

Scenario 1: The searcher thinks that *T* is related to his medical situation according to the information provided in P_T^o. In this case, the information in the symptom set, the answer set, and the other-inputs set has led him to find *T* and is no longer important. The searcher prefers to read more Web pages about certain aspects of *T*. To serve this purpose, P_a provides multiple *aspect elements*, one for each of the following aspects: (1) symptom and sign, (2) diagnosis, exam, and test, (3) treatment, (4) cause and trigger, (5) risk factor, (6) complication, (7) medication, (8) surgery, (9) prognosis (expectation), (10) rehabilitation, recovery, self-care, and home treatment, (11) complementary and alternative medicine, (12) prevention, and (13) resource, support, living with, and management.

Scenario 2: From the information provided in P_T^o, the searcher cannot determine whether *T* is related to his medical situation. In this case, the information in the symptom set, the answer set, and the other-inputs set is still important. The searcher prefers to read more Web pages about *T*. To serve this purpose, P_a provides two aspect elements, one for each of the following aspects: (1) general information, and (2) other information.

As mentioned in Section II-B, for each aspect *A* of disease *T*, iMed provides an overview Web page $P_{A,T}^o$. To obtain $P_{A,T}^o$, we form a query $Q_{A,\,T}$. The first result Web page retrieved by $Q_{A,\,T}$ is used as $P_{A,T}^o$. The snippet of $P_{A,T}^o$ is obtained using both $Q_{A,\,T}$ and standard passage retrieval techniques [16].

We form query $Q_{A,\,T}$ in the following way. For each aspect element prepared for Scenario 1, $Q_{A,\,T}$ includes the following two sets of information:

Set 1 (disease set): *T*.

Set 2 (aspect set): *A*.

Both sets have the same weight 1. The keywords of the aspect set are obtained from the name of *A*. We use a weight-constrained method similar to that described in Section III-B-2 to strike a balance between the contribution from the disease set and the contribution from the aspect set in $Q_{A,\,T}$. For the general information aspect element prepared for Scenario 2, $Q_{A,\,T}$ includes the keywords of *T*. For the other information aspect element prepared for Scenario 2, we use as $Q_{A,\,T}$ the same Q_T that is used to retrieve the overview Web page P_T^o for disease *T* at the first level of the search result hierarchy (see Section III-B-2). In this case, since the first result Web page retrieved by Q_T has been used as P_T^o, the second result Web page retrieved by Q_T is used as the overview Web page $P_{A,T}^o$ for *A* of *T*.

4) *Third Level of the Hierarchy:* In this section, we discuss how to construct the third level of the search result hierarchy. At this level, we use a query $Q_{A,\,T}$ to retrieve Web pages for aspect *A* of disease *T*. This $Q_{A,\,T}$ is the same one used to retrieve the overview Web page $P_{A,T}^o$ for *A* of *T* at the second level of the search result hierarchy (see Section III-B-3). Since the first result Web page P_1 retrieved by $Q_{A,\,T}$ has been used as $P_{A,T}^o$, P_1 is excluded at the third level of the search result hierarchy. (If *A* is the other information aspect, the first two result Web pages retrieved by $Q_{A,\,T}$ have been used at the first two levels of the search result hierarchy and hence are excluded at the third level of the search result hierarchy.) Each retrieved Web page's snippet is obtained using both $Q_{A,\,T}$ and standard passage retrieval techniques [16].

C. Step 3: Suggesting Medical Phrases

In this section, we describe how to suggest related medical phrases. We focus on the first level of the answer interface. The other levels can be handled in a similar way. In general, good medical WSEs should automatically suggest diversified, related medical phrases [12], [20], [38] to help searchers quickly digest search results and refine their inputs (e.g., the keywords in the "other inputs" text area of iMed). These suggested medical phrases should be ordered by their relevance to the searcher's inputs.

iMed extracts and ranks medical phrases based on multiple sources: the MeSH ontology [22], the collection *C* of crawled Web pages, and the formed queries. The suggestion process contains two sub-steps and recommends medical phrases in the MeSH ontology.

1) *Sub-step 1*: In the first sub-step, iMed selects *V* candidate medical phrases from the returned top-*J* overview Web pages, where the default values of *V* and *J* are 60 and 20, respectively. The suggested medical phrases need to be both relevant and diverse in order to provide the greatest convenience to the searcher [20]. Intuitively, to ensure that a medical phrase *M* is relevant, *M* better appears in one of the returned top Web pages with a large tf×idf value. More related medical phrases should be suggested for those formed queries with larger weights. To ensure enough diversity in the list of suggested medical phrases, a single Web page should not contribute too many medical phrases to that list.

We use a continuous discounting method to achieve these goals. For each medical phrase *M* in a Web page *P* that is retrieved for a formed query *Q*, we compute a weighted tf×idf value $w_{M,\,P}$ that is the product of *Q*'s weight and *M*'s original tf×idf value in *P*. Each time a medical phrase is selected from *P*, a discount is given to the weighted tf×idf values of the remaining medical phrases in *P*. As a result, the more medical phrases have already been selected from *P*, the less likely the remaining medical phrases in *P* will be selected in the future. Also, more related medical phrases are likely to be suggested for those queries with larger weights. The concrete method is as follows.

For each of the returned top-*J* Web pages, we find all its medical phrases and compute their tf×idf values using the Okapi formula [20]. We obtain a list L_t of triplets (medical phrase *M*, Web page *P*, weighted tf×idf value $w_{M,\,P}$), and select *V* distinct medical phrases from L_t to form a candidate set *S*. This is done in *V* passes. In each pass, a medical phrase M' with the largest weighted tf×idf value is selected from L_t. Then all the triplets with the same medical phrase M' are dropped from L_t, as we are only interested in distinct medical phrases. For all the remaining medical phrases in the Web page where M' comes from, their weighted tf×idf values are discounted by a factor *d* whose default value is 0.9.

2) *Sub-step 2*: In the second sub-step, we rank all the medical phrases in the candidate set *S* and present them to the searcher. A simple method, which we call the *weighted tf×idf method*, is to rank all these medical phrases in the order that they are generated in the first sub-step. As explained in [20] and we will show in Section IV-C, the quality of the resulting order is often unsatisfactory. A better method, which we call the *weighted relevance score method*, is to consider the different weights of all the formed queries and to rank all these medical phrases in descending order of their relevance scores for the queries. In computing relevance scores, we address the terminological discrepancy between queries and medical phrases by "translating" both of them into plain English description using the representative page technique [20]. The concrete method is as follows.

For each medical phrase in the MeSH ontology, iMed retrieves offline the top-ranked *r* Web pages as *M*'s representative Web pages, where *r* is a constant. For each formed query *Q*, we use the top-ranked *s* Web pages retrieved for *Q* as *Q*'s representative Web pages. Here *s* is a constant.

Consider a medical phrase $M \in S$ coming from a Web page *P* that is retrieved for a formed query *Q*. The relevance score between *M* and *Q* is computed as a weighted average of the cosine similarities [3] of their representative Web pages:

$$score_M = \sum_{i=1}^{r} \sum_{j=1}^{s} \text{cosine\_similarity}_{R_i, P_j} / (i \times j) \cdot$$

Here, the weight for the *i*-th (1≤*i*≤*r*) representative Web page R_i of *M* is 1/*i*, and the weight for the *j*-th (1≤*j*≤*s*) representative Web page P_j of *Q* is 1/*j*. Each medical phrase corresponds to a single query and has only one relevance score. In cosine similarity computation, we use the techniques in [30] to reduce noise: (1) In R_i of *M* (or P_j of *Q*), we only consider those terms whose distances from a term in *M* (or *Q*) are no more than W_1 terms, where W_1 is a predetermined window size; (2) Among those terms, we only consider the top W_2 ones with the largest tf×idf values, where W_2 is a predetermined constant.

In general, we should rank higher the medical phrases that have larger relevance scores and correspond to those formed queries with larger weights. For diversity purpose, we also prefer that highly ranked medical phrases do not all correspond to the same few queries with the largest weights. To achieve these goals, we use another continuous discounting method to select all the medical phrases in the candidate set *S* one by one and rank them in that order. This method is similar to the one used in the first sub-step with the following three differences: (1) For each medical phrase *M* corresponding to a query *Q*, we compute a weighted relevance score that is the product of *M*'s relevance score and *Q*'s weight; (2) Formed queries replace Web pages; (3) Weighted relevance scores replace weighted tf×idf values.

IV. EXPERIMENTAL RESULTS

We conducted experiments under a wide range of medical scenarios to demonstrate the effectiveness of our proposed techniques.

A. Setup

iMed is a vertical WSE that crawls Web pages from a few selected, high-quality medical Web sites instead of the entire Web. In our experiments, we crawled 22GB of Web pages from WebMD [35], Healthline [12], and Merck [23], three of the most popular medical Web sites. These Web sites cover the entire medical domain fairly comprehensively and include information on various topics such as symptoms, diseases, drugs, and treatments.

We compared iMed with two state-of-the-art medical WSEs: Google Health [10] and Healthline [12]. We used both real medical case records from the Family Medicine Online Database (FMOD) [8] and USMLE medical exam questions [34]. Correct diagnoses are available for both of them and serve as the ground truth for our evaluation. USMLE stands for the United States Medical Licensing Examination, whose exam question format is similar to the format of actual, well-documented medical case records. Physicians have to pass this exam to obtain their licenses for practicing medicine. In our tests, each exam question is treated as a medical case. FMOD was developed by the College of Medicine of the Pennsylvania State University for educating medical students. The FMOD records document patients' medical situations in great detail using mostly layman terms and can be easily understood by ordinary people.

> Mrs. Brown is a 67-year-old woman with a two-day history of swelling and pain in her right lower leg. The pain worsens with walking. She remembers first noticing right calf pain when arising from bed yesterday morning. She noticed the right lower leg was swollen and red. She denies any traumatic event or recent strenuous activity. In fact, she just returned from her winter stay in Florida, and the long drive gave her legs a needed rest. The leg pain worsens with walking, especially pushing off with her toes. The pain is relieved with rest and elevation but her calf continues to hurt. She describes the pain to be most severe inside her calf. She has not had this leg swelling and pain previously and has never had leg edema. She denies fever or night sweats. She cannot recall any reason for her skin to be infected. She recalls no bites or scrapes to her calf area. Four months ago, Mrs. Brown had a breast mass removed that was found to be malignant. A local surgeon is following her and told her the nodes were cancer free and the tumor was completely removed ...

Fig. 6. An example medical case record (www.hmc.psu.edu/ume/fcmonline /case4/index.htm).

We randomly selected 30 medical cases from the FMOD records and the USMLE exam questions. One such medical case is shown in Fig. 6. Since both USMLE and FMOD cover almost every aspect of medical practice, our random samples have a broad coverage of medical topics.

In our experiments, a user has up to 60 minutes to perform iterative search for each medical case. At the end of the search process, the user can list up to three diseases that he thinks best match the medical case's situation description. If any of these diseases is among the correct diagnoses accompanying the data set, the search is considered successful. We allow users to search for a relatively long time, because medical information searchers care about their health and often spend hours on searching. We allow users to list multiple diseases as their findings, because even doctors sometimes cannot make precise diagnosis without lab test results.

Ten colleagues served as assessors and users. None of them has formal medical training. For a medical case, each user randomly selected one of the three medical WSEs (iMed, Google Health, or Healthline) with equal probability to perform search. Our experiments were performed on a computer with two 3GHz processors, 2GB memory, and one 111GB disk.

Similar to the TREC interactive track [33], we use two sets of measures as the performance metrics for medical WSEs: one set is objective while the other set is subjective. The objective performance measures include the success rate, the number of search iterations, the number of search result Web pages viewed, and the time spent on the search process. The subjective performance measures include the users' perceptions of ease of using the system, ease of understanding the system, usefulness of the search results, and overall satisfaction with the system. For iMed, both the average usefulness of the overview Web pages for the top 10 diseases (for determining whether these diseases are related to the medical case's situation description) and the average usefulness of the top 10 suggested medical phrases at the first level of the search result hierarchy are also included. All these subjective performance measures are on a 7-point scale, with 1=low and 7=high [33]. They were obtained from a brief questionnaire that users filled out after using the systems. For each objective or subjective performance measure, we average it over all the 30 medical cases and all the users, and report both its mean and its standard deviation when appropriate. We used ANOVA [2] as the significance test.

B. An Example

TABLE I. SOME RETURNED RELEVANT WEB PAGES.

rank	URL	topic
1	www.webmd.com/hw/parenting_news/hw81306.asp@printing=true	upper respiratory system infections
2	www.webmd.com/hw/health_guide_atoz/tm1440.asp@printing=true	lung cancer
3	www.webmd.com/hw/heartburn/hw99227.asp	gastroesophageal reflux disease
4	www.webmd.com/asthma/guide/asthma-smoking	quit smoking
5	www.webmd.com/hw/vision/aa118910.asp@printing=true	toxic fumes
7	www.healthline.com/galecontent/hay-fever	hay fever

To give the reader a feeling of the contents returned by iMed, we present detailed results of the returned Web pages and the suggested medical phrases for a typical query scenario that corresponds to choosing “little or no sputum” and “no dyspnea” for the symptom cough (see Fig. 3). Table I shows some relevant Web pages returned at the first level of the search result hierarchy. The suggested relevant medical phrases include silicosis (rank 1), smoking cessation (rank 2), pneumoconiosis (rank 3), esophagitis (rank 4), respiratory system (rank 5), and bacterial infections (rank 7). For a query scenario Q_s, iMed generally can find several relevant Web pages and medical phrases describing multiple topics related to Q_s.

C. Overall Results

In this section, we present the overall experimental results. iMed is efficient at performing medical search. For all the 30 medical cases, the average time taken by iMed to generate each part of the search result hierarchy is less than two seconds. iMed’s is much more effective than the other two medical WSEs in finding the correct diagnosis, where most of the user’s time is spent on reading the search result Web pages. The objective performance measures in Table II show that compared to the other two medical WSEs, iMed makes the user find results in fewer iterations, view fewer search result Web pages, spend less time on the search process, and achieve a higher success rate. All these differences are statistically significant.

TABLE II. OBJECTIVE PERFORMANCE MEASURES (* MEANS SIGNIFICANT at <0.05 LEVEL COMPARED to IMED).

mean (standard deviation)	iMed	Healthline	Google Health
success rate	30% (12%)	23%* (9%)	21%* (10%)
number of iterations	3.9 (1.2)	5.9* (1.5)	6.1* (1.4)
number of search result Web pages viewed	14 (6.2)	20* (7)	21* (7.2)
time (minutes)	31 (11)	41* (12)	43* (12)

Table III shows the subjective performance measures. All the users are familiar with the traditional keyword query interface and the sequential order presentation of search results. It took these users a while to become accustomed to navigating the search result hierarchy in iMed’s answer interface. As a result, users think that the traditional WSE user interface is slightly easier to understand than iMed’s user interface, while the difference is not statistically significant. Nevertheless, once users understand iMed’s user interface, they can use it without difficulty. iMed’s answer interface has explicitly marked medical meanings and organizes together all the search results on the same topic or aspect so that users can find them easily. Users are also accustomed to using questionnaires in daily life. Consequently, users think that iMed’s user interface is easier to use than the traditional WSE user interface. Overall, users think that iMed produces more useful search results and is more satisfactory than the other two medical WSEs. These differences are statistically significant.

TABLE III. SUBJECTIVE PERFORMANCE MEASURES (* MEANS SIGNIFICANT at <0.05 LEVEL COMPARED to IMED).

mean (standard deviation)	iMed	Healthline	Google Health
ease of using	5.7 (1.2)	4.9* (1.0)	4.9* (1.0)
ease of understanding	5.6 (1.1)	5.8 (1.1)	5.8 (1.1)
usefulness	5.2 (0.9)	4.3* (1.0)	4.2* (0.9)
satisfaction	5.0 (1.0)	4.2* (0.9)	4.0* (0.9)

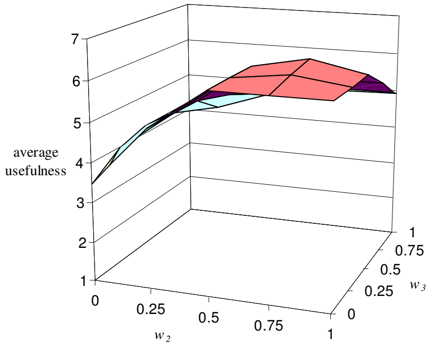


Fig. 7. Average usefulness of the overview Web pages for the top 10 diseases.

For iMed, Fig. 7 shows the impacts of the weights w_2 and w_3 (see Section III-B-2) on the average usefulness of the overview Web pages for the top 10 diseases. When either w_2 or w_3 is small, information about the searcher’s medical situation is not fully incorporated into the formed queries. When both w_2 and w_3 are large, the keywords of the answer set and the symptom set overwhelm the keywords of the disease set. Consequently, the retrieved overview Web pages are only marginally relevant to the corresponding diseases.

In the weighted relevance score method, s (or r) is the number of representative Web pages for each formed query (or medical phrase). The default values of s and r are both 1. We varied s from 1 to 4 and r from 1 to 3. Fig. 8 shows the impacts of s and r on the average usefulness of the top 10 suggested medical phrases at the first level of iMed’s search result hierarchy. The horizontal dotted line represents the average usefulness when the weighted relevance score method is not used and the suggested medical phrases are ranked using the weighted tf×idf method described in Section III-C-2. In general, for a query (medical phrase), the higher-ranked representative Web pages are more relevant than the lower-ranked representative Web pages. Hence, the weighted average usefulness score decreases as s or r increases. To achieve good performance, it is best to set $s=1$ and $r=1$. Furthermore, if the weighted relevance score method is not used, the quality of suggested medical phrases degrades by 20%.

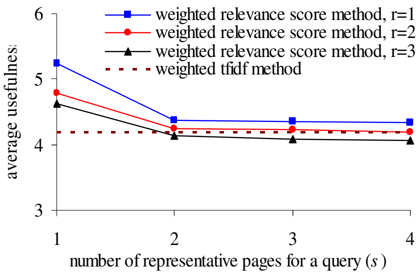


Fig. 8. Average usefulness of the top 10 suggested medical phrases.

V. CONCLUSION

This paper presents iMed, the first intelligent medical Web search engine that extensively uses medical knowledge and questionnaire to facilitate ordinary Internet users to search for medical information. The design of iMed takes into consideration the unique requirements of medical search. Instead of asking searchers to form queries themselves, iMed uses a questionnaire-based query interface to guide searchers to provide the most important information about their situations. iMed requires no special user training, forms queries automatically, structures all the search results into a multi-level hierarchy that has explicitly marked medical meanings, and suggests related medical phrases. These features are attractive to ordinary Internet users who have little medical background. Our experiments with a wide range of medical scenarios demonstrate that iMed greatly improves user satisfaction by performing medical search effectively and efficiently. For future work, we are interested in using our techniques to build intelligent search engines for other domains (e.g., product search).

ACKNOWLEDGEMENT

We thank Haiyan Chen, Jiuxing Liu, Linda Schumer, Chunqiang Tang, Selena Thomas, Ying-li Tian, Jing Wang, Leiping Wang, and Hong Xu for helpful discussions.

REFERENCES

[1] *American Medical Association Family Medical Guide*, 4th ed., John Wiley & Sons, 2004.

[2] P.J. Bickel and K.A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics, Vol. 1*, Prentice Hall, 2001.

[3] R.A. Baeza-Yates and B.A. Ribeiro-Neto. *Modern Information Retrieval*, ACM Press/Addison-Wesley, 1999.

[4] R.D. Collins, *Algorithmic Diagnosis of Symptoms and Signs: Cost-Effective Approach*, Lippincott Williams & Wilkins, 2002.

[5] J.G. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proc. SIGIR'98*, 1998, pp. 335-336.

[6] (2008) Conditions by Incidence. [Online]. Available: http://www.wrongdiagnosis.com/lists/incid.htm.

[7] (2007) Data & Statistics, Centers for Disease Control and Prevention. [Online]. Available: http://www.cdc.gov/DataStatistics.

[8] (2008) Family Medicine Online homepage. [Online]. Available: http://www.hmc.psu.edu/ume/fcmonline/index.htm.

[9] (2006) 'Googling' Aids Difficult Diagnoses. [Online]. Available: http://www.e-health-insider.com/news/item.cfm?ID=2258.

[10] (2008) Google Health homepage. [Online]. Available: http://www.google.com/Top/Health.

[11] D.R. Goldmann, *American College of Physicians Complete Home Medical Guide*, DK Publishing, 2003.

[12] (2008) Healthline homepage. [Online]. Available: http://www.healthline.com.

[13] P.M. Healey and E.J. Jacobson, *Common Medical Diagnoses: An Algorithmic Approach*, 2nd ed., W.B. Saunders, 1994.

[14] A.L. Komaroff, *Harvard Medical School Family Health Guide*, Free Press, 2004.

[15] D.L. Kasper, E. Braunwald, and A. Fauci *et al.*, *Harrison's Principles of Internal Medicine*, 16th ed., McGraw-Hill Professional, 2004.

[16] M. Kaszkiel and J. Zobel, "Passage retrieval revisited," in *Proc. SIGIR'97*, 1997, pp. 178-185.

[17] G. Luo and C. Tang, "On iterative intelligent medical search," in *Proc. SIGIR'08*, 2008, pp. 3-10.

[18] G. Luo and C. Tang, "Challenging issues in iterative intelligent medical search," in *Proc. ICPR'08*, 2008.

[19] G. Luo, "Intelligent output interface for intelligent medical search engine," in *Proc. AAAI'08*, 2008, pp. 1201-1206.

[20] G. Luo, C. Tang, and H. Yang *et al.*, "MedSearch: a specialized search engine for medical information retrieval," in *Proc. CIKM'08*, 2008.

[21] (2008) Medstory homepage. [Online]. Available: http://www.medstory.com.

[22] (2008) MeSH homepage. [Online]. Available: http://www.nlm.nih.gov/mesh/meshhome.html.

[23] (2008) Merck Manual Home Edition homepage. [Online]. Available: http://www.merck.com/mmhe/index.html.

[24] (2007) New Healthline symptom search dramatically improves one of the most popular online health research activities. [Online]. Available: http://www.healthline.com/corporate/news /healthline_announces_symptom_search.html.

[25] F. Radlinski and S. Dumais, "Improving personalized Web search using result diversification," in *Proc. SIGIR'06*, 2006, pp. 691-692.

[26] P. Ramnarayan, A. Tomlinson, and G. Kulkarni *et al.*, "A novel diagnostic aid (ISABEL): development and preliminary evaluation of clinical performance," in *Proc. Medinfo'04*, 2004, pp. 1091-1095.

[27] S.E. Robertson, S. Walker, and M. Hancock-Beaulieu, "Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive," in *Proc. TREC'98*, 1998, pp. 199-210.

[28] R.H. Seller, *Differential Diagnosis of Common Complaints*, 4th ed., W.B. Saunders, 2000.

[29] (2008) SearchMedica - The GPs search engine. [Online]. Available: http://www.searchmedica.co.uk/searchmedica/EUIHomeAction.do.

[30] M. Sahami and T.D. Heilman, "A Web-based kernel function for measuring the similarity of short text snippets," in *Proc. WWW'06*, 2006, pp. 377-386.

[31] C. Sherman. (2005) Curing medical information disorder. [Online]. Available: http://searchenginewatch.com /showPage.html?page=3556491.

[32] (2008) SMART Stopword List. [Online]. Available: http://www.lextek.com/manuals/onix/stopwords2.html.

[33] (2008) TREC Interactive Track homepage. [Online]. Available: http://trec.nist.gov/data/interactive.html.

[34] (2008) USMLE homepage. [Online]. Available: http://www.usmle.org.

[35] (2008) WebMD homepage. [Online]. Available: http://www.webmd.com.

[36] J. Williams, "When expert systems are wrong," in *Proc. ACM SIGBDP Conf. on Trends and Directions in Expert Systems 1990*, pp. 661-669.

[37] L. Yi, B. Liu, and X. Li, "Eliminating noisy information in Web pages for data mining," in *Proc. KDD'03*, 2003, pp. 296-305.

[38] Q.T. Zeng, J. Crowell, and R.M. Plovnick *et al.*, "Assisting consumer health information retrieval with query recommendations," *JAMIA* 13(1), pp. 80-90, 2006.

[39] Q.T. Zeng and T. Tse, "Exploring and developing consumer health vocabularies," *JAMIA* 13(1), pp. 24-29, 2006.