Improving Result Adaptation through 2-step Retrieval

Meike Reichle and Kerstin Bach

Intelligent Information Systems Lab
University of Hildesheim
Marienburger Platz 22, 31141 Hildesheim, Germany
{reichle|bach}@iis.uni-hildesheim.de,

Abstract. In this paper we present the retrieval and adaptation mechanisms used in our information system on travel medicine, docQuery. The retrieval method's main feature is an overall improved accuracy of retrieval results' similarities through a more diverse distribution of similarities over the retrieved result sets. Its underlying idea is the execution of several consecutive retrievals on one case base, where attributes from the cases resulting from the first query are used to refine a subsequent query in order to yield better results than the first retrieval. The refined result sets narrow down the search space for cases to be used in result adaptation, which improves adaptation quality. The mechanisms are implemented in the docQuery information system on travel medicine.

1 Introduction

Intelligent information systems provide a technology for covering even complex topics in a comprehensive but flexible way. Realising such systems requires high quality data sources, knowledge models, and maintenance techniques. To achieve this, knowledge has to be acquired, analysed, stored, and retrieved, which challenges a knowledge-based system and is crucial for its continuous existence over a longer period of time. Case-Based Reasoning (CBR) is a methodology that has proven most effective for knowledge storage, retrieval and adaptation in all kinds of intelligent information systems [1].

In this paper we present the 2-step retrieval mechanism, a retrieval mechanism for CBR systems that works in an iterative way, executing two consecutive retrieval steps on the same case base, using information gained from the results of the first retrieval step in order to refine the second one. This consecutive retrieval strategy leads to an overall improved accuracy of retrieval results' similarities through a more diverse distribution of similarities over the retrieved result set.

Our first application of the 2-step retrieval mechanism is the docQuery project [2], an intelligent information system on travel medicine that is being developed in a joint project by the Intelligent Information Systems Lab and mediScon worldwide. docQuery is built using the SEASALT architecture [3], which is a first instantiation of the CoMES approach [4].

The paper is structured as follows: Section 2 presents the docQuery project, the application domain travel medicine and its particular challenges with regard to knowledge-based systems. Section 3 presents the actual 2-step retrieval algorithm, with an illustrated example. Section 4 finally presents an evaluation of the 2-step retrieval algorithm based on its application in the docQuery project, describing the evaluation's setup in subsection 4.1 and its results in subsection 4.2 followed by a description of related approaches in section 5. The paper closes with a conclusion and an outlook on future work in section 6.

2 Travel Medicine as an Application Domain: the docQuery System

Travel medicine is an interdisciplinary speciality concerned with the prevention, management and research of health problems associated with travel and covers all medical aspects a traveller has to take care of before, during and after a journey. For that reason it covers many medical areas and combines them. Furthermore, information about the destination, the activities planned and additional conditions have to be considered when giving medical advise to a traveller. Travel medicine starts when a person moves from one place to another by any kind of transportation and stops after returning home healthy. In case a traveller gets sick after a journey a travel medicine consultation might also be required.

The research project presented in this paper is supported by mediScon worldwide, a Germany based company with a team of physicians specialized on travel medicine and TEMOS[1], a tele-medical project of the Institute of Aerospace Medicine at the German Aerospace Center DLR[2]. Together we will develop *docQuery*, an intelligent information system on travel medicine which provides relevant information for each traveller for their individual journey. First of all we will focus on prevention work, followed by information provision during a journey and information for diseased returnees.

Since common sources on the World Wide Web are not authorized by physicians and/or experts, we aim at providing reliable information for everybody. In preparation for a healthy journey it is important to get a high quality and reliable answer on travel medicine issues which both laymen and experts should be able to use. Based on the SEASALT architecture [3], we propose building a web community which gives information to travellers and physicians (non-experts in the field of travel medicine) by experts on travel medicine. docQuery will provide an opportunity to share information and ensure a high information quality because it is maintained by experts. Furthermore it will rise to the challenge of advancing the community alongside their users. Travellers and experts can visit the website to get the detailed information they need for their journey. A traveller will give docQuery the key data on their journey (like travel period,

---

[1] TEMOS means TElemedicine for a MObile Society, see http://www.temos-network.org

[2] http://www.dlr.de/me/

destination, age(s) of traveller(s), activities, etc.) and docQuery will prepare an information leaflet the traveller can take to his or her general practitioner to discuss the planned journey. The leaflet will contain all the information needed to be prepared and provide detailed information if it is required. In the event that docQuery cannot answer the traveller's question, the request will be sent to experts who will answer it. The computation of the answer follows the steps a physician would take during a consultation. Since travel medicine touches on different topics such as geographic information, diseases, medicaments, activities etc. We developed a modularised knowledge base, with a case base for each respective topic. These case bases are queried and their results are then combined, observing the constraints given by the user and domain itself (e. g. medical preconditions or medicines that cannot be taken in combination).

Modularising case bases into subdomains instead of simply partitioning them into smaller ones with the same case format, has several advantages. Firstly the individual case bases are easier to maintain with regard to the correctness of their contents, since they represent a more simple knowledge domain. Also, breaking up the rather complex domain of travel medical advisories into more simple subdomains that are than recombined as needed, gives the whole information system more flexibility. Providing the appropriate combination rules exist, the contents of the individual case bases can also be combined into cases that have not yet been presented to the system, as long as they adhere to the respective combination rules.

Further, not all knowledge domains that are included in an information system on travel medicine require the same type of maintenance and are subject to the same amount of change over time. While it is for instance no problem to keep a rather simple domain such as countries and regions as minimal and consistent as possible, the domain of travel related diseases is better served by including as many cases as possible, even if some of them are very similar. By splitting the knowledge domain into these smaller subdomains, each of them can be maintained in a way that is best suited for the respective subdomain.

Since the topic of this paper is the retrieval on one individual case base, we only gave a short overview of the docQuery System and its modularised case bases in this section. More on modularised case bases, their maintenance, and the combination of their results can be found in [5].

3   2-Step Retrieval

When dealing with a topic like travel medicine we cannot assume that all users ask complete and/or correct questions and like Weibelzahl [6] we enrich the user's query enhancing it with additional information from the case base.

Our initial retrieval is based on the geographic position of a country and because of the fact that the earth is divided in a manageable amount of countries, which are completely covered in our case base's similarity measure in the form of a geographic taxonomy, we can rely that every requested country can be retrieved. However, we cannot be sure that we will have (complete) information on

that country. Also, due to the nature of our domain, travel medicine, geographic proximity is not sufficient to find feasible adaptation candidates to complete the retrieved country's information – also occurring diseases have to be noticed. In our retrieval mechanism we thus start by requesting the destination country, this step is followed by an enhanced query including additional information about the initial country's diseases. The second retrieval's results with the highest similarity will be the adaptation candidates we take into account. In the current approach we randomly pick one of the cases with the highest similarity as adaptation source. In the future we will add maintenance information to the cases to be able to compute the most updated or most recently maintained adaptation candidate.

The approach presented here concentrates on interdependent attributes that are not completely given for every single case. We will show how we can narrow the result set by retrieving reliable information snippets in order to adapt them to create a (complete) response.

Assuming that we have a traveller planning a journey, the retrieval will start based on the destination region. We know that our model contains all countries of the world, so the retrieval algorithm will be able to find the appropriate destination. But due to many changes of disease outbreaks we have to provide up to date country and regional information. Therefore we do not only retrieve the country we have searched for, we also include in our result set countries with a similar *structure* considering travel medical aspects. To realize this we also use information about vaccinations that can be divided in three categories:

1. Obligatory Vaccinations: Those vaccinations are required in order to be allowed enter a country.
2. Standard Vaccinations: Those vaccinations are required if one is travelling to a certain country - although they are not a regulation.
3. Risk Vaccinations: Those vaccinations are required for people with an enfeebled immune system such as pregnant women, children, elderly people, or those who suffer from different kinds of (chronic) diseases and require a higher protection provided by a vaccination.

Additionally our system will give information about diseases that can be contracted during a journey. According to [7] those can be divided in the following categories of diseases: vectors (In medicine a vector is a carrier of infections, diseases, etc. because it carries for example the parasitic agent i.e. in malaria a mosquito serves as the vector), person-to-person contact, ingestion of food and water, bites and stings, and water/environmental contact.

Currently we do have information about vaccination advices, but we do not have complete information about diseases contracted during a journey, because they rely until a certain point on up-to-date information. Nevertheless we will provide this kind of information to the users of docQuery and since we do have similarity models for each type of disease we will adapt the information from similar countries.

In order to ensure that our system adapts correct data we will use the 2-Step-Retrieval to reduce the amount of cases we can adapt from. In the first

step we will only do the retrieval based on our geographic taxonomy, then we narrow the set of retrieved cases by adding vaccination information to a second query. The taxonomy includes 228 countries and islands arranged by continents, subcontinents (e.g. Western Europe), regions (e.g. Iberian Peninsula) followed by the country. A generalization step leads to a value of 0.5 and specialization to a value of 0.8.

When performing a standard 1-step retrieval on Laos, a whole of 10 countries in it's geographic proximity have a similarity of 40%. When performing a 2-step retrieval, the distribution of similarities is much more diverse, as illustrated in figure 1.
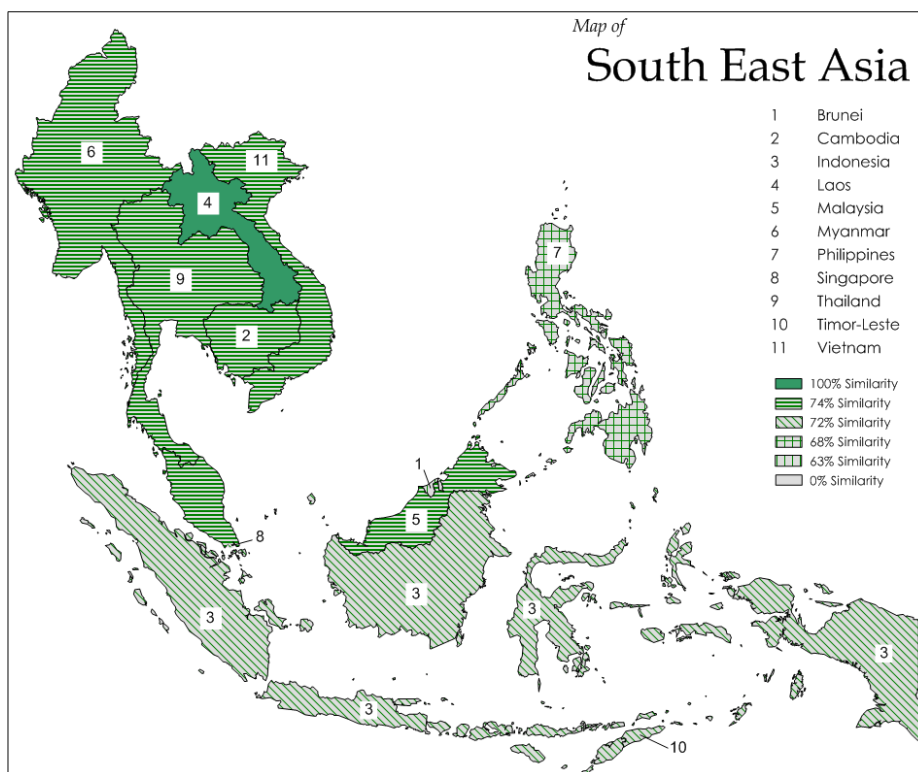
Fig. 1. 2-Step-Retrieval for the example of Laos

Laos does not yet have complete information on the diseases contracted there, so these attributes have to be filled using adaptation from another, similar case. To complete the disease information we now have to choose one country to adapt from. Using 1-step retrieval there are 10 adaptation candidates to chose from,

some of which, such as e. g. the Philippines are in fact quite different from Laos with regard to travel medicine.

To ensure that the cases we adapt from are more similar to the requested country, we now also consider vaccination information in our request, using it in a second retrieval step as illustrated in figure 2. Since we need information on countries with a similar *disease structure* in order to be able to find a country profile with an appropriate amount of information, even if the destination originally given by the user does not offer those, we use 2-step retrieval. An example query (again using the country, this time plus the vaccination information of the originally retrieved country) for the second retrieval step would be: *"Laos, Yellow fever, Diphtheria, Hepatitis A, Measles, Tetanus, Cholera, Hepatitis B, Japanese Encephalitis, Rabies, Typhoid fever"*.

After the second retrieval step, the countries situated around Laos now have more differentiated similarities and offer a higher amount of information considering their *disease structure*. As can be seen in figure 1, this time only the countries near Laos as well as Malaysia are returned with the same similarity to Laos, reducing the number of adaptation candidates to 5. If we are taking one of those countries into account the likelihood of retrieving a valid result set will be highly increased.
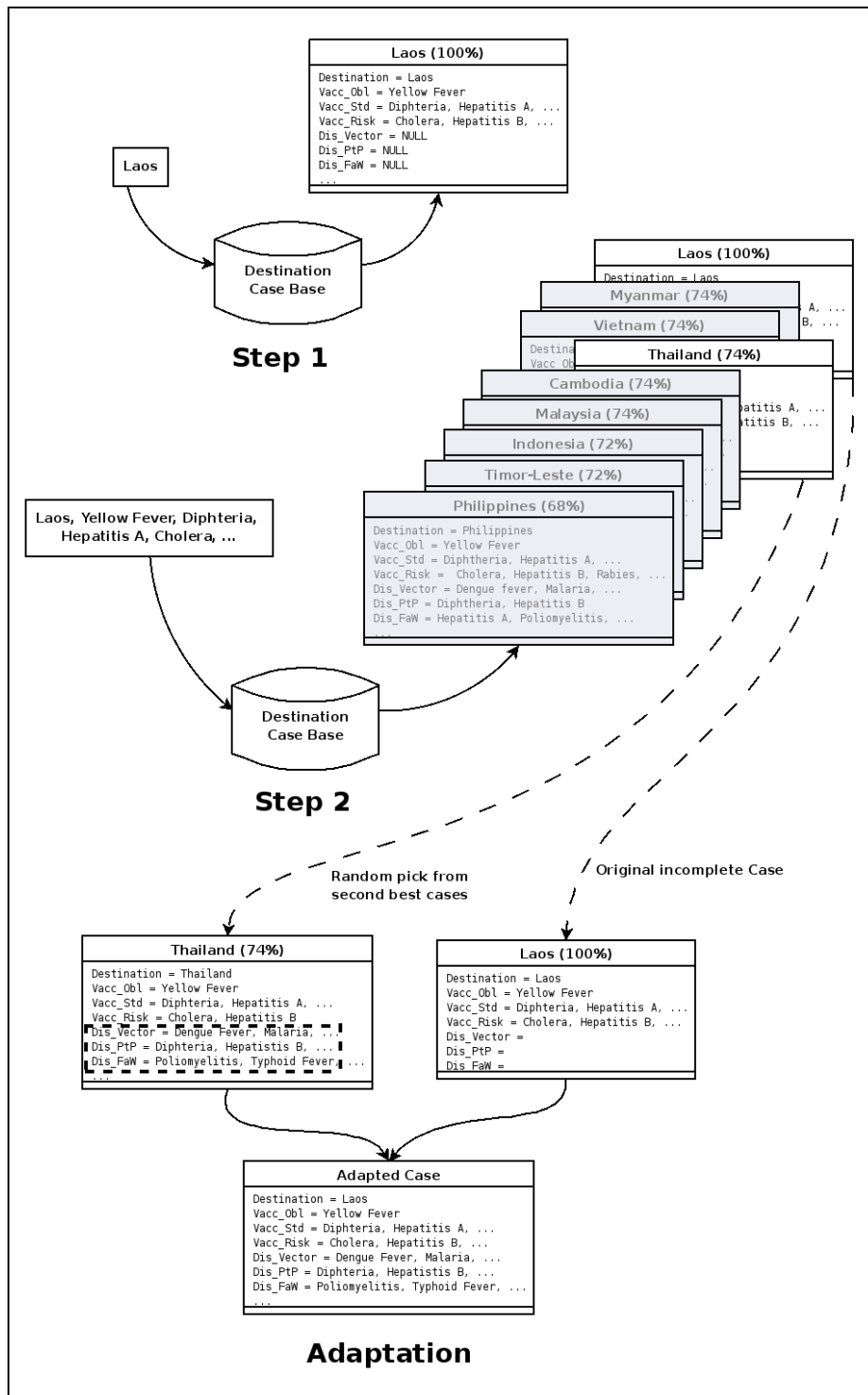
4 Evaluation: 2-Step-Retrieval

Following the example given in section 3 we will now present the evaluation of our approach in the travel medicine domain. Therefore we will present how the 2-Step-Retrieval affects the retrieved result sets and we illustrate its advantage in comparison to a straightforward 1-step retrieval approach. In the architecture of docQuery, the case bases *Destination*, *Associated Information*, and *Activity* are adequate to do 2-step retrieval, since our other case bases contain health critical information that require a more strict retrieval.

4.1 Experimental Setup

The *Destination* case base covers country characteristics that are used to prepare an information leaflet for a traveller. It contains vaccination requirements and vaccination-preventable infectious diseases, pre-travel information on different kinds of diseases that might occur in a certain country or region, as well as hygiene and prevention advise.

The experimental data contain a case base covering all countries in the world and the vaccination information abroad we have to consider preparing information leaflets for travellers. To carry out the experiment we took a controlled sample of 18 countries of East and South East Asia, representative with respect to country borders, coasts, islands and climatic conditions and manually filled in the data on transmittable diseases, so that we have 18 cases with complete information. The sample comprises all countries of East and South East Asia that

Fig. 2. 2-step retrieval and adaptation using Laos as the query and Thailand as the randomly picked adaptation candidate.

ensures that we are able to find neighbouring countries, non-neighbouring countries, and countries associated to different nodes (e.g. China which is situated in Eastern Asia and Thailand which is in South East Asia).

We carried out our evaluation as a leave-one-out experiment. For each country in the case base we did the following steps:

1. Remove diseases information from country.
2. Do a one-step retrieval using the country name.
3. Do a two-step retrieval using the country name and vaccination information as additional information.
4. Do an adaptation with each respective result set.
5. Compare the set of diseases obtained from the two respective adaptation candidates to the original set of diseases.

For the experiment we use the following weights:

$$sim = 6 \times [Region] + 4 \times [Vacc\_Risk] + 3 \times [Vacc\_Std] + 2 \times [Vacc\_Obl] \quad (1)$$

In our similarity measure "*Region*" is weighted times 6 because it is the most diverse and reliable fact, and because we expect that travellers know their destination, but not the diseases. "*Risk people vaccination*" information are weighted times 4 because vaccination advices depend on each traveller's profile and in particular his or her disease history in combination with chronic illness(es). Also, this attribute differentiates countries from each other. "*Standard vaccination*" advices are weighted times 3 because they describe the disease structure from the medical point of view and allow a general classification. "*Obligatory vaccination*" is weighted times 2 because obligatory vaccination requirements depend on the geographic region as well as on official orders.

For each country we did 5 runs[3] and compared the result of the 1-step-retrieval and the 2-step-retrieval to the expected result. First we compared the number of adaptation candidates and then the resulting adaptation quality, checking if all expected disease were found, if one or more diseases were missed (*false negatives*), or if incorrect extra diseases were added to the case (*false positives*).

4.2 Results of the Experiment

Figure 3 shows how the number of adaptation dropped significantly in 90% of our test cases. These numbers illustrate that result sets are indeed more diverse when enriching the queries with extra information. Moreover not only the number of adaptation candidates change, also the adaptation candidates differ between 1-step and 2-step retrieval. For example the retrieved countries for Japan: In the 1-step-retrieval North Korea, Mongolia, Taiwan, China, as well as Macau are returned, but the adaptation candidates for the 2-step retrieval do not include Mongolia, but Hong Kong, because of the fact that the disease structure of Hong

---

[3] We chose to perform several runs, since the randomising element in the final choice of the adaptation candidate can yield different results for the same query.

Kong is much more similar to Japan than Mongolia. In all test cases at least one country has been dropped out of the adaptation candidates, but on average 3.3 countries were not taken into account in the 2-step retrieval. In 39% of the cases an adaptation candidate has been add.
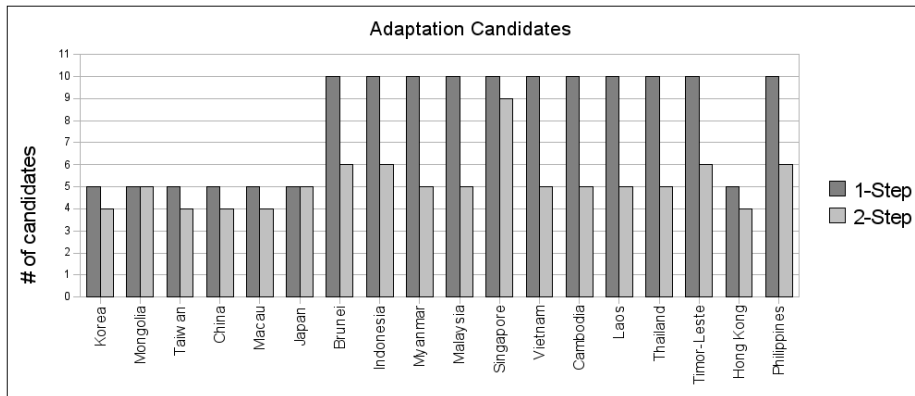
Fig. 3. Comparison of the number of adaptation candidates in 1- and 2-step retrieval

In the next step we investigated whether the adaptation candidates remaining after the 2-step retrieval were in fact the better ones, that is, if their adaptation results were better than the adaptation results from the candidates resulting from one-step retrieval. The results of the adaptations can be seen in Figure 4.
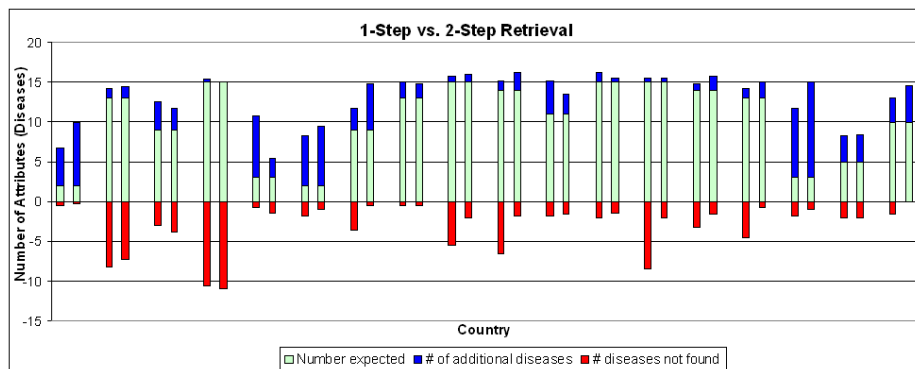
Fig. 4. Result Comparison of the 1-Step vs. the 2-step Retrieval

Each column pair represents the aggregated results of one country – the first column shows the adaptation results of the 1-step-retrieval and the second of the

2-step-retrieval. The y-axis shows the number of correct diseases (*true positives*, positive scale, light-coloured) and extra (*false positives*, positive scale, dark-coloured) diseases found as well as missed diseases (*false negatives*, negative scale, dark-coloured). The result sets used for the evaluation result from requesting the following countries: Korea, Mongolia, Taiwan, China, Macau, Japan, Brunei, Indonesia, Myanmar, Malaysia, Singapore, Vietnam, Cambodia, Laos, Thailand, Timor-Leste, Hong Kong, and the Philippines.

In total we did 90 single-case requests and after the 1-step-retrieval 62% of the adapted cases contained all of the expected diseases. Applying the 2-step retrieval to the same cases 76% of the adapted cases contained all expected diseases. Although both retrieval variants also return false positives in most of the tests, the solutions of the 2-step retrieval are generally more reliable, especially according to false negatives. The 2-step retrieval performed significantly better than the 1-step retrieval with regard to false negatives as can be seen in Figure 5.
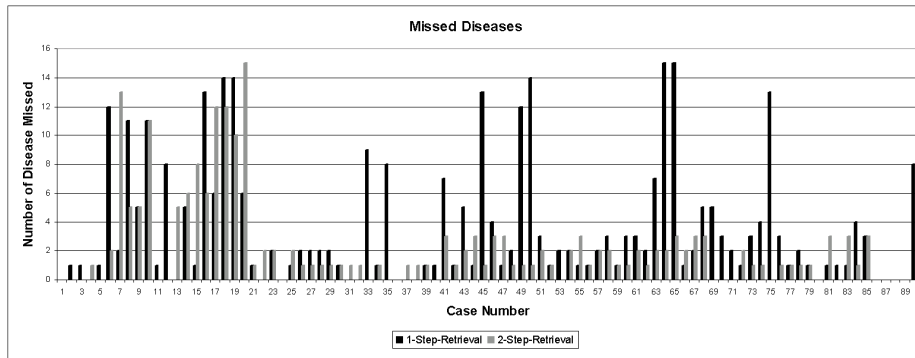

Fig. 5. Unaggregated comparison of missed diseases (false negatives).

In summary the experiment shows that the 2-step retrieval provides more robust results and especially in the field of travel medicine a query can be significantly enhanced by adding more information on the destination, because, for example, disease and vaccination advice can only be provided by an expert, not by the traveller. Furthermore the information stored in the case bases can be used to create more refined queries.

5 Related Work

The idea of using and combining information from different cases has also been discussed in [8], in which Redmond describes how snippets of different cases are combined to receive a solution for a given problem. In comparison to our approach, Redmond uses similar case representations from which he extracts

parts of cases in order to combine them, but in our approach we take the whole retrieved case as a snippet of our solution. Nevertheless those two approaches have in common that each snippet has to match the other snippets and limits solutions that go along with it.

A similar approach has been presented in [9, 10] in which the incremental CBR (I-CBR) mechanism for diagnosis has been introduced. The I-CBR separates information in between ”‘free”’ and ”‘expensive”’ features and starts the first retrieval steps based on the free features before the user is asked to give information about expensive features to narrow the set of information. In comparison with this approach we have a different point of view. Our system already holds the user's information and we do not necessarily narrow the result set, but we use the 2-step retrieval to tighten the set of candidates we derive information from to adapt single information. Another approach on how I-CBR can influence the result sets has been presented in [11], but in comparison to our approach Jurisica et. al. did not receive additional information from exiting case, they used query series and user interaction instead.

In [6, 12] Weibelzahl uses a travel domain consisting of two case bases with different knowledge models. The first case base, called customer case base, holds information on the customers' needs and desires which are mapped to attributes describing products provided in the second case base. In the first step the query containing the user's expectation on their vacation is analysed to set relevant attributes creating a request which can be sent to the product case base regarding the users' expectations. The second request contains especially those product attributes the user would not request on their own, but help to find an appropriate solution in the product case base.

6   Conclusion and Outlook

In this paper we have presented a retrieval mechanism that enhances a query with information in order to receive a more diverse result set. Using 2-step retrieval provides us with robust results to dispatch the subsequent retrieval and result combination. The 2-step retrieval algorithm presented in this paper exemplifies how the retrieval strategy can be implemented in a CBR system. Further on we use this approach to combine and adapt parts of cases and attributes of different case bases, because we expect that our information obtained of the travel medicine community will be incomplete. Also we suppose that taking more attributes into account might help the algorithm to receive even more diverse result sets.

As a next step we will enable our system to combine the retrieval results of cases retrieved of modularised heterogeneous case bases in order to create a whole individual information leaflet for travellers containing information on activities, diseases, medication, etc. Hence, we will implement a multi-agent system centred around a coordination agent (or broker agent) combining retrieval results and ensuring complete information regarding given constraints.

Another aspect of our future work is generalising the 2-step retrieval algorithm and evaluating whether the algorithm can by applied to other case bases and domains as well, or if this only works for our specific domain. Also, we have to figure out if the algorithm of using retrieval results for refining queries can be applied to other case bases in the travel medicine application domain as well as in other application domains.

Integrating the 2-step retrieval algorithm in SEASALT puts forward our idea of using knowledge lines for building CoMES upon existing knowledge sources.

References

1. Bergmann, R., Althoff, K.D., Breen, S., Göker, M.H., Manago, M., Traphöner, R., Wess, S.: Selected Applications of the Structural Case-Based Reasoning Approach. In: Developing Industrial Case-Based Reasoning Applications: The INRECA-Methodology. Volume 1612 of Lecture Notes in Computer Science. Springer (2003) 35–70
2. Bach, K.: docquery - a medical information system for travellers. Internal project report (September 2007)
3. Bach, K., Reichle, M., Althoff, K.D.: A domain independent system architecture for sharing experience. In: Proceedings of LWA 2007, Workshop Wissens- und Erfahrungsmanagement. (September 2007) 296–303
4. Althoff, K.D., Bach, K., Deutsch, J.O., Hanft, A., Mänz, J., Müller, T., Newo, R., Reichle, M., Schaaf, M., Weis, K.H.: Collaborative multi-expert-systems – realizing knowlegde-product-lines with case factories and distributed learning systems. In Baumeister, J., Seipel, D., eds.: Workshop Proceedings on the 3rd Workshop on Knowledge Engineering and Software Engineering (KESE 2007), Osnabrück (September 2007)
5. Althoff, K.D., Reichle, M., Bach, K., Hanft, A., Newo, R.: Agent based maintenance for modularised case bases in collaborative multi-expert systems. In: Proceedings of AI2007, 12th UK Workshop on Case-Based Reasoning. (December 2007) 7–18
6. Weibelzahl, S.: Conception, implementation, and evaluation of a case based learning system for sales support in the internet. Master's thesis, Universität Trier (1999)
7. The International Society of Travel Medicine: The body of knowledge for the practice of travel medicine (2003)
8. Redmond, M.: Distributed cases for case-based reasoning: Facilitating use of multiple cases. In: AAAI. (1990) 304–309
9. Cunningham, P., Bonzano, A., Smyth, B.: An incremental case retrieval mechanism for diagnosis (1995)
10. Cunningham, P., Smyth, B., Bonzano, A.: An incremental retrieval mechanism for casebased electronic fault diagnosis (1998)
11. Jurisica, I., Glasgow, J., Mylopoulos, J.: Incremental iterative retrieval and browsingfor efficient conversational cbr systems. Applied Intelligence **12**(3) (2000) 251–268
12. Weibelzahl, S., Weber, G.: Benutzermodellierung von Kundenwünschen durch Fallbasiertes Schliessen. In Jörding, T., ed.: Adaptivität und Benutzermodellierung in interaktiven Softwaresystemen, ABIS-99, Magdeburg (1999) 295–300