Similarity of Medical Cases in Health Care Using Cosine Similarity and Ontology. Shahina Begum, Mobyen Uddin Ahmed, Peter Funk, Ning Xiong, Bo von Schéele (PBMStressMedicine AB), in the proceedings of the 5th Workshop on CBR in the Health Sciences, p 263-272, Springer LNCS, Belfast, Northern Ireland, August, 2007

Similarity of Medical Cases in Health Care Using Cosine Similarity and Ontology

Shahina Begum, Mobyen Uddin Ahmed, Peter Funk, Ning Xiong, Bo von Schéele

Mälardalen University, Department of Computer Science and Electronics
PO Box 883 SE-721 23, Västerås, Sweden
{firstname.lastname}@mdh.se

Abstract. The increasing use of digital patient records in hospital saves time and reduces risks of wrong treatments caused by lack of information. Digital patient records also enable efficient spread and transfer of experience gained from diagnosis and treatment of individual patient which is now-a-days mostly manual (speaking with colleagues) and rarely aided by computerized system. Most of the content in patient records is semi-structured textual information. In this paper, we propose a hybrid textual case-based reasoning system promoting experience reuse. This is derived from structured or unstructured patient records, case-based reasoning and similarity measurement based on cosine similarity metric improved by a domain specific ontology and the nearest neighbor method. As a result, hospital staffs can learn not only new cases but also add comments to existing cases and thus enables prototypical cases.

1 Introduction

Patient records are increasingly stored digitally in a computer readable form. Patient records can be accessed even from other hospitals in emergency situation. This opens up new possibilities for experience reuse. The use of medical cases is established both in education and in every day work by medical staff. In the Care-Partner [1] project, it has been proven that a multimodal reasoning framework is able to identify relevant cases if they are represented in a structured way. A case retrieval framework is described in [12] where the authors applied textual CBR approach to acquire and elicit knowledge from structured documents. In the legal domain, a textual case-based reasoning (TCBR) system [4] using information gain algorithm along with cosine similarity was used for classification. For retrieving textual cases, feature vector generalization was used to form structural cases [14] where it captures semantic relationships by the way of association. In [9], a vector space model-based retrieval system using cosine similarity and manual weighting for full text document search in MEDLINE is presented (our system uses a domain specific ontology instead of manual weighting). In [2], an incident report retrieval system combines traditional CBR with Cosine similarity function to find similarities and patterns.

A modified cosine matching function was used in the electromechanical domain to contrast cases in [3] where it showed better performance in retrieval compared to Nearest Neighbor. However, the cases may be ill structured [13] or having structures that do not match between cases, especially when digitalizing past cases or they may contain terminology that does not in accordance to the clinical standard.

To enable similarity matching both on structured and less structured cases containing text, we propose the combination of cosine similarity with synonyms and ontology. If cases are well structured, the cosine similarity can be used on the structured parts individually provided they both have the same structure (some parts may be more important than others, e.g. the symptom description may be more important than the treatment section in a patient record when looking for a similar patient case). If cases are lacking of structures, e.g., older cases, the cosine similarity will still be able to identify relevant cases. Since natural language systems at present are unable to fully understand the meaning of the text, textual matching mostly benefits from combining different similarity measurements as they often complement each other. If the patient's record also contains traditional features, e.g. age, sex, prescribed medication doses etc, the textual similarity and the similarity in features need to be weighted together before the final rank. The cases are communicated in an appropriate way to the clinician. By knowing the user's current context (at desktop or in Operation Theater) and the user's profile (e.g. experience in the particular case), the system can interact more efficiently with the user. Cases may also be hypothetical, e.g., hypothetical cases containing some missed symptoms and tests which results in wrong treatment with severe consequences. In similarity with such "negative" cases, it is essential to alert clinical staff.

2 System Overview

An illustration of how the system may work in a medical context is given in Fig. 1 where the user starts with a new medical record. Ideally this patient record (case) is given in a structured way (structured text) and with explicit features. When hospital staff type in a new patient record, the system may immediately start to search for similar cases. If it is a common case (e.g. a flue) then the system may not initiate an interaction with the user, but if there is a similarity with e.g. a meningitis case that can not be ruled out, the medical staff may be alerted. The proposed system should be context aware and interact with the user in an appropriate way, i.e., being in an operating theater or in the office may require a very different dialogue, indicated by (2) in Fig. 1. The system may also allow the user to make more elaborated searches for similar cases based on the patient record and some additional information (text and/or features labeled as (1) in Fig. 1). The system facilitates the sharing of experience stored both locally and globally by enabling clinical staff to comment cases (labeled as (3) in Fig. 1). The system also promotes collaboration by providing contact details to an expert or colleague that has encountered a patient with similar symptoms recently; the system may even inform if the colleague is at work or not.

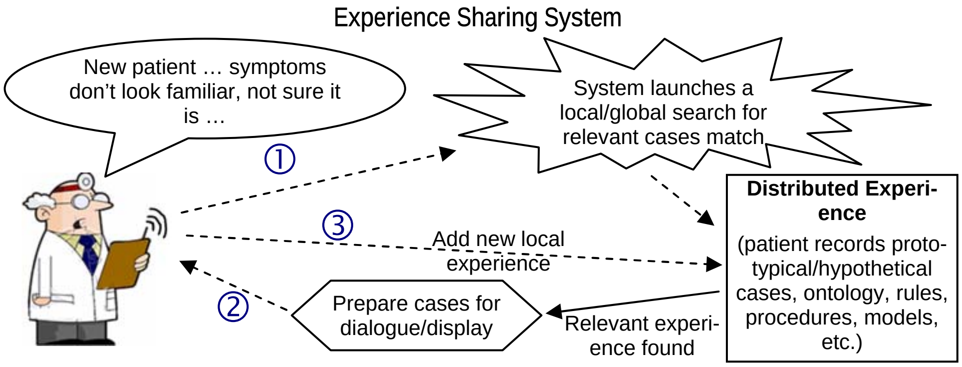

Experience Sharing System
New patient … symptoms don't look familiar, not sure it is …
①
System launches a local/global search for relevant cases match
Distributed Experience
(patient records proto-typical/hypothetical cases, ontology, rules, procedures, models, etc.)
③
Add new local experience
②
Prepare cases for dialogue/display
Relevant experience found

Fig. 1. An overview of the system's functionality

To exemplify how the proposed communication system in fig. 1 would behave in a real-world application, two scenarios are described in table 1. Each health care environment such as a department or hospital may contain a local experience sharing system, but the system may extend its search in other areas, e.g. collaborating hospitals, public and commercial experience libraries etc. Both examples involve advanced search while basic search is performed automatically when new patient records are entered into the system.

Table 1. Example scenarios in health care

Example1. Disease diagnosis by sharing experience:	*Example2. Experiences achieved for deciding tests leading to efficient diagnosis:*
A clinician studies a patient's record and makes a hypothesis on what disease the patient may have. The clinician uses the patient record together with the hypothesis as seed to the search. The system performs a search locally for the similar cases but the local system fails to identify a similar and relevant case. The system then launches a global request for similar cases. The global search is performed by other experience sharing systems located in the collaborating hospitals and a number of similar cases are identified and securely transferred to the clinician. The clinician then uses the provided information to decide for a number of additional tests and also communicates with one of the authors of one of the cases using skype. Some revision is made for the case and after verification the clinician reports this information to the local experience sharing system and save it as a new case.	A clinician makes a hypothesis based on a patient's history and symptoms and suggests a set of tests to confirm his diagnosis. But the clinical tests do not confirm the hypothesis and at the same time both the local and global experience sharing system failed to find any suitable solution. So the clinician suggests another set of tests to discover the true cause and this additional tests report shows the indication of a disease that is not correspond to the initial hypothesis. From this incident the clinician has gained new knowledge. The clinician also makes a personal reflection to the case anticipating that it will save future clinicians from making the same initial mistake and spending substantial costs on a number of negative tests.

3 Representation of experiences as cases

Experience can be represented as a contextualized piece of knowledge in a case. The case typically consists of a problem specification and solution where we can store most types of data such as textual values (e.g. names, addresses), numeric values (e.g. cost, ages, blood pressure) and multimedia features (e.g. photographs, sound, and video). Table 1 gives a picture of an example case from Parkinson's disease adapted from [10].

Table 2. Example of a case in health science

Case type: Advise alternate medication

Case name: Parkinson predominant with akinesia and rigidity.

History: A sixty-five years old male person suffered from Parkinson's disease for two years predominated with akinesia and rigidity. He has minimal asymmetrical tremor and used to take Immediate-Release Sinemet four times a day. The patient appears to be a therapeutic responder to L-dopa and he has no definite off time.

Symptoms: The patient has slight morning akinesia and problem in doing daily activates. But he does not have motor fluctuations and has no high dyskinesia.

Features: The patient is a male and he is 65 years old. He has been taking Sinemet Immediate-Release between 400 and 600 mg per day at 4 times since last 2 years.

Action taken: Modified the medications to improve the activities of daily living by adding a half of a tablet of 25/100 to either two or four of the doses of regular Sinemet. By observing the patient it was found that his daily living activities improves using this medication. Another way of generally helping the activities of daily living would be to add a dopamine agonist.

Requirements and tools: Necessary to have a proper follow up by the clinician and may need hospitalization

Outcome: The solution is acceptable

Suggestions and summary: A dopamine agonist probably would be well tolerated since the disease has not been present for many. This would help improve a patient who was maybe sub-optimally treated generally. The adding of a dopamine agonist would also bring about the theoretical consideration of neuroprotection. Since this patient is a nonfluctuator, a drug such as Tasmar could be used three times a day. Tasmar in nonfluctuators also reduced the likelihood of developing motor fluctuations as compared to placebo patients.

The experience adding interface allows flexibility in mapping experiences into cases by adapting different situations and problems and user can also define their own experience structures by themselves. A problem description can be represented both in textual format as well as with a number of features. In this system, users can enter text to describe symptoms, diagnosis, cause analysis and history; it has also the option to choose the outcome and indicates the success rate and upload files as attachment. The format to represent experiences embeds the same underlying standard structure that is generally used to record medical cases and well understood and accepted throughout the users.

4 Retrieval of similar cases

A CBR system generally includes the essential steps such as retrieval, reuse, revise, and retain. The retrieval step is the first step where the aim is to find the most similar cases which have potential to be reused. The procedure of case retrieval begins with identifying the most important features and uses them in identifying cases to reuse. For the textual cases, the *tf-idf* (term frequency–inverse document frequency) [6] weighting scheme is used in the vector space model [7] together with cosine similarity to determine the similarity between two cases [11]. Additional domain information often improves results, i.e., a list of words and their synonyms or dictionaries provides comparable words [8] [5] and relationships within the words using class and subclass. Our proposed system uses domain specific ontology that represents specific knowledge, i.e., relation between words. The different steps in retrieval of similar case(s) in the system are shown in fig 2.

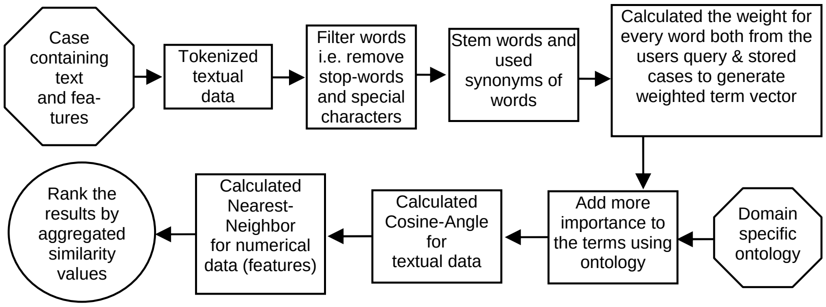

Case containing text and features
Tokenized textual data
Filter words i.e. remove stop-words and special characters
Stem words and used synonyms of words
Calculated the weight for every word both from the users query & stored cases to generate weighted term vector
Domain specific ontology
Add more importance to the terms using ontology
Calculated Cosine-Angle for textual data
Calculated Nearest-Neighbor for numerical data (features)
Rank the results by aggregated similarity values

Fig. 2. Different steps for case retrieval

The text tokenizer algorithm decomposes the whole textual information into sentences, and then into individual words. A filtering step is required to improve retrieval effectiveness due to the huge amount of words. The following three steps are used to extract the important textual features:

1. Remove the stop-words and special characters blacklist both from the users' query and patients' record.
2. A list of synonyms of the words is used to reduce the number of terms and Porter stemming algorithm [7] helps stemming the words that provide the ways of finding morphological variants of search term. After calculating the weight for each word, these words are represented as terms in a vector space.
3. Improve the importance assessments for candidate terms before measuring the cosine similarity value for the textual information between the stored case and user's query case by using domain specific ontology.

Different features values along with local weights were used to find similarity between the features of stored cases and the features of new case where the Nearest-Neighbor algorithm works perfectly. For presenting a sorted list of results, we aggregate all local similarity values both for a user's query and stored case, which is described in the following section.

4.1 Term frequency and weighting

The weighting terms ($W_{i,j}$) method calculates the weight of each term or word from the stored cases and the inputted user's query to perform further matching. The general equation for $W_{i,j}$ is shown in equation (1). Where, $W_{i,j}$ is the weight of term T_j in the case C_i, $tf_{i,j}$ is the frequency of term T_j in the case C_i and idf_j is the inverse case frequency where N is the number of cases in the database and df_i is the number of cases where term T_j occurs at least once.

$$W_{i,j} = tf_{i,j} * idf_j = tf_{i,j} * \log_2 (\frac{N}{df_i}) \qquad \textbf{(1)}$$

The new case is processed according to the vector space model and stored in a separate table. First, an index of the terms from the case collection is constructed and the frequency of the terms ($tf_{i,j}$) appearing in each case (C_i) and new query case (Q) is counted. Then, the case frequency (df_i) from the collection of cases and the inverse case frequency (idf_j) are calculated and finally, the $tf_{i,j}* idf_j$ product gives the weight for each term.

4.2 Enhanced term vector using the domain specific ontology

Each word of a case can be treated as a term and it is easy to calculate the weight of each term for every case where terms of each case are satisfied with other case by exact matching or by synonym or having a co-occurrence. However, still some words or terms which have a complex relationship (for example, the term fluctuation and L-dopa) can be defined by ontology and the weight of those terms can be increased automatically for that case using the domain specific ontology defined by the expert. We can enhance the weight of the vector terms for each case based on the following conditions:

Condition-1: If a term T_f in the case is related to a term T_o in the ontology but the term T_o does not exist in the case, then the term T_o can be added as a *new* term with the same importance as the weight of the source term i.e. the score of *tf-idf.*

Condition-2: If a term T_f in the case is related to a term T_o in the ontology and also the term T_o exists in the case, then the strength of relationship between the term T_f and T_o can be added to the original weight (i.e. score of *tf-idf*) of those terms.

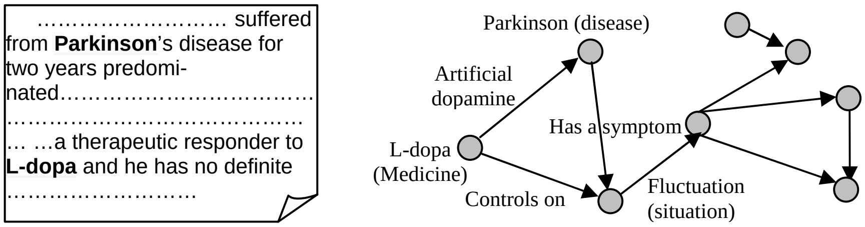
Condition-3: If more than one term in a case are related to a term T_o in the ontology, then those terms of that case will get more importance by adding their relationship strength to their original weight (i.e. score of *tf-idf*).

Condition-4: If a term T_f in a case is related to more than one term in the ontology then the normalized strength of their relationship can be added to the original weight of source term T_f.

An example is shown in fig.3 to show how the ontology helps to improve the weight vector.

From Fig. 3, we see that “L-dopa” is a term that appears both in the case text and in ontology, and has a relation with the term “fluctuation”. But the term “fluctuation” does not exist in the case text, so the term “fluctuation” is important for this case and can be added according to condition 1. Again the terms “Parkinson” and “L-dopa” both already exist in the case text and have a relation in the ontology, so the value of their strength of relationship for those two terms (“Parkinson” and “L-dopa”) will increase their importance (condition 2). Terms “Parkinson” and “L-dopa” are related to another term “fluctuation” in the ontology so the term “fluctuation” will get more importance according to condition 3.

Condition 4 is the vice versa of condition 3. Thus the terms will get importance assessments depending on the ontology & hence allow an improved similarity measurement.



……………………… suffered from **Parkinson**'s disease for two years predominated………………………………………………………………………… … …a therapeutic responder to **L-dopa** and he has no definite ………………………

Fig. 3. Increasing the weight of a term using ontology

4.3 Similarity matching for textual part

To find the textual similarity between a stored case vector C_i and a new case query vector Q we apply cosine similarity function [7] [3] [14] for the textual information. This ratio is defined as the cosine of the angle between the vectors, with values between 0 and 1 and can be calculated by equation 2.

$$Cos\theta_{C_i} = Sim(Q,C_i) = \frac{Q \bullet C_i}{\|Q\|\|C_i\|} = \frac{\sum_j w_{q,j} w_{i,j}}{\sqrt{\sum_j w_{q,j}^2}\sqrt{\sum_j w_{i,j}^2}} \qquad \textbf{(2)}$$

Where $Cos\theta_{C_i}$ is the *cosine of the angle* between a stored case and a query case which is defined as the similarity function *Sim(Q, C_i)*. The dot product is calculated between the stored case and the query case by $Q.C_i$ where zero products are ignored; then vector $\|Q\|\|C_i\|$ lengths are calculated for a stored case and a query case where $w_{i,j}$ and $w_{q,j}$ are weights calculated through equation 1(zero terms also ignored).

4.4 Similarity matching for structured features

The Nearest-Neighbor (NN) method is a common matching technique in most CBR systems. The similarity between user's query case and stored cases are calculated using the NN method where the Euclidian distance measurement is used between the numerical features. The semantics of similarity for a symbolic feature is usually defined by domain experts in the form of a numeric matrix quantifying the degrees of similarity for every pair of symbolic values associated with that feature. The general equation of the NN method is described in equation 3

$$Similarity\ (T,S) = \sum_{f=1}^{n} W_f * sim(T_f, S_f) \tag{3}$$

In equation 3, *Similarity (T, S)* calculates global similarity for all the structured features, *T* is the query/current case, *S* is the stored case, W_f is the normalized weight defined by each local weight (given by the user) divided by sum of local weights of individual features *f* , *n* is the number of the features in each case, *f* is the individual feature from 1 to *n*, and *sim* is the local similarity function for feature in case *T* and *S*.

$$sim(T_f, S_f) = 1 - \frac{abs(T_f - S_f)}{Max(T_f, S_f) - Min(T_f, S_f)} \tag{4}$$

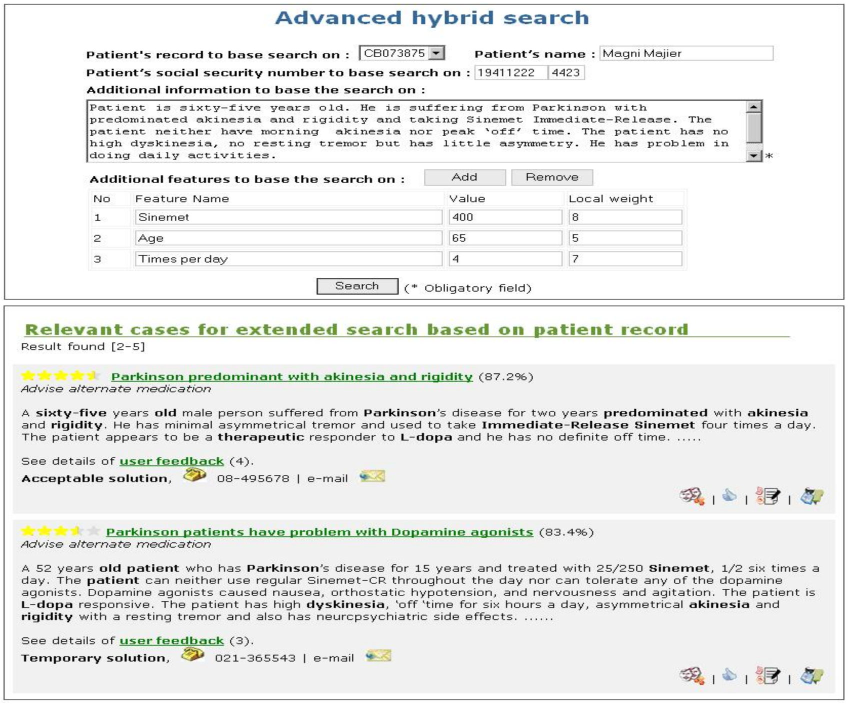
$sim(T_f, S_f)$ in equation (4) represents local similarity function for numerical features, function *abs* is used to get the absolute values and *Max* and *Min* of the features values are derived from the whole case base and user query.

5. Results

The result from the system formulates a ranked list of similar cases based on the aggregation of the local similarity values where all the cases are listed according to percentage and 100% means the perfect match. Acceptable similar cases are presented along with their outcomes (i.e. acceptable or temporary solution etc.) and success rates from users' feedback. The detail problem, solution and its related parts can be shown according to the defined security level in each case author's profile. For instance, if a matching case is unrestricted or local, the solution is shown to others. Otherwise the user has to contact the owner of the case to get access to the solution. A screen shot for a hybrid advanced search for similar experiences is shown in Fig 4.

The users can enter a textual query and also several features for searching related experiences and the most relevant experiences are presented with a case title, type, description of history and symptoms along with the scores (similarity values). The user is able to set several search criteria in the advanced search option, for example, keywords matching, synonyms, stemming, ontology etc. The system can also detect spelling errors through the word dictionary (WordNet).

In the result, the system also provides additional interactive options (using 4 different symbols below each case) such as, a user can make further matching which means it can show the same type of other cases if required, can rate each case, can enable a user to report on the experience after reviewing the selected experience or a user can see others comments on the selected experience and at the same time can provide his/her own comments. In the proposed system, the medical staff can provide his/her feedback or comments on that experience(s) and rank it on how much the experience has been matched with his/her current experience which could be a valuable information for the future users to avoid expensive mistakes due to lack of experiences.



Advanced hybrid search

Patient's record to base search on : CB073875

Patient's name : Magni Majier

Patient's social security number to base search on : 19411222 4423

Additional information to base the search on :

Patient is sixty-five years old. He is suffering from Parkinson with predominated akinesia and rigidity and taking Sinemet Immediate-Release. The patient neither have morning akinesia nor peak 'off' time. The patient has no high dyskinesia, no resting tremor but has little asymmetry. He has problem in doing daily activities. *

Additional features to base the search on :

No	Feature Name	Value	Local weight
1	Sinemet	400	8
2	Age	65	5
3	Times per day	4	7

(* Obligatory field)

Relevant cases for extended search based on patient record

Result found [2-5]

Parkinson predominant with akinesia and rigidity (87.2%)

Advise alternate medication

A **sixty-five** years **old** male person suffered from **Parkinson**'s disease for two years **predominated** with **akinesia** and **rigidity**. He has minimal asymmetrical tremor and used to take **Immediate-Release Sinemet** four times a day. The patient appears to be a **therapeutic** responder to **L-dopa** and he has no definite off time. .....

See details of user feedback (4).

Acceptable solution, ☎ 08-495678 | e-mail

Parkinson patients have problem with Dopamine agonists (83.4%)

Advise alternate medication

A 52 years **old patient** who has **Parkinson**'s disease for 15 years and treated with 25/250 **Sinemet**, 1/2 six times a day. The **patient** can neither use regular Sinemet-CR throughout the day nor can tolerate any of the dopamine agonists. Dopamine agonists caused nausea, orthostatic hypotension, and nervousness and agitation. The patient is **L-dopa** responsive. The patient has high **dyskinesia**, 'off 'time for six hours a day, asymmetrical **akinesia** and **rigidity** with a resting tremor and also has neurcpsychiatric side effects. ......

See details of user feedback (3).

Temporary solution, ☎ 021-365543 | e-mail

Fig. 4. List of relevant cases based on an advanced hybrid search

6. Summary and conclusions

Experience sharing and reuse is becoming increasingly beneficial in health care as it extends the knowledge and capability of clinicians in disease diagnosis and treatments. This paper develops a case-based reasoning system capable of representing and handling experiences of clinicians in patient record cases containing both structured and unstructured data. The case consists of both explicit features and their corresponding values of either symbolic or numerical nature and descriptions of perceived symptoms, feelings of patients, as well as contextual information usually represented in a natural language. In the paper, we propose an approach for enhancing the performance of textual matching and retrieval by incorporating domain knowledge with the help of ontology. It has also been shown that a mixture of structured and textual data can be manipulated in a united framework with CBR for experience sharing and transfer in health care applications.

The significance of accommodating textual data in medical CBR research is: 1) it presents a useful attempt to capture perception-based experiences coming from human observations and feelings. Unlike measurement-based experiences, human perceptions are usually expressed in an informal and natural language format, but they are proved important for diagnosis as well in addition to objective sensor readings. 2) the work would enable better contextual awareness for decision support on diagnosis and treatment plans. Frequent contextual information is conveyed in relevant notes or reports; hence hybridization with textual data enables new possibilities of utilizing contextual awareness in a medical CBR system leading to more reliable and effective experience reuse.

References

1. Bichindaritz I., Siadak M. F, Jocom J., Moinpour C., Kansu E., Donaldson G, Bush N, Chapko M., Bradshaw J. M., Sullivan K. M. CARE-PARTNER: a Computerized Knowledge-Support System for Stem-Cell Post-Transplant Long-Term Follow-Up on the World-Wide-Web", Journal of American Medical Informatics Association (JAMIA): 386-390, Suppl. for AMIA'98 Annual Symposium. (1998)
2. Cassidy, D., Carthy, J., Drummond, A., Dunnion, J, Sheppard, J.: The Use of Data Mining in the Design and Implementation of an Incident Report Retrieval System, Proceedings of the Systems and Infor-mation Engineering Design Symposium (2003)
3. Gupta K.M., Montazemi A.R.: Empirical Evaluation of Retrieval in Case-Based Reasoning Systems Using Modified Cosine Matching Function, IEEE transactions on systems, man, and cybernetics—part a: systems and humans, vol. 27, no. 5 (1997).
4. Proctor J. M., Waldstein I., Weber R.: Identifying Facts for TCBR. 6th International Conference on Case-Based Reasoning, Workshop Proceedings. Stefanie Brüninghaus (Ed.) Chicago, IL, USA, August 23-26, (2005) 150-159
5. Recio J. A., Díaz-Agudo B, Gómez-Martín M.A. and Wiratunga N.: Extending jCOLIBRI for textual CBR. In Procs. Of 6th International Conference on CBR, volume 3620 of LNCS, Springer –Verlang, (2005) 421-435.
6. Salton G. and C. Buckley: Term Weighting Approaches in Automatic Text Retrieval, Technical Report. UMI Order Number: TR87-881., Cornell University (1987).
7. Salton G., A. Wong and C. S. Yang.: A Vector Space Model for Automatic Indexing, Communications of the ACM, vol.18, nr. 11, (1975) 613–620.
8. Scott S. and Matwin S.: Text Classification Using WordNet Hypernyms, Use of Word-Net in Natural Language Processing Systems (1998).
9. Shin K., and Sang-Yong H.: Improving Information Retrieval in MEDLINE by Modu-lating MeSH Term Weights, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 978-3-540-22564-5, Volume 3136, (2004) 388-394.
10. Silver D.E, Case Studies in Parkinson's Disease, http://www.pdasd.org/site/index.asp?page=88475&DL=7246, Last referred on May (2007)
11. Weber, R., Ashley K. D., and Brüninghaus S. B.: Textual case-based reasoning, The Knowledge Engineering Review, Vol. 00:0, 1–00., Cambridge University Press, Printed in UK (2005).
12. Weber, R.; Aha, D., Sandhu, N., and Munoz-Avila H.: A Textual Case-Based Reasoning Framework for Knowledge Management Application, In Proceedings of 9th GWCBR, (2001) 40-50.
13. Wilson D. C., and Bradshaw, S.: CBR Textuality, Expert Update, 3(1). (2000) 28-37.
14. Wiratung_a N., Koychev I. and Massie S.: Feature Selection and Generalisation for Retrieval of Textual Cases in Proceedings of the 7th European Conference on Case-Based Reasoning, Springer-Verlag, (2004) 806–820.