Methodology for Classifying and Indexing Case-Based Reasoning Systems in the Health Sciences

Isabelle Bichindaritz and John C. Reed Jr.
University of Washington Tacoma
1900 Commerce Street
Tacoma, Washington 98402, USA
ibichind@u.washington.edu

Abstract

As the amount of information available to researchers grows at an increasing rate, it becomes much more difficult to find relevant resources. An approach taken by several authoritative bodies, such as the Association for Computing Machinery and the U.S. National Library of Medicine, is the introduction of a classification scheme. However, even the most modern schemes are not capable of adequately distinguishing one research paper from another, due mainly to their broad generality. This paper describes a methodology for building a much narrower, specialized classification scheme focused on the area of Cased-Based Reasoning in the Health Sciences. It is derived from thorough analysis of the field, but with a framework that can be adapted to other areas. Using a tiered approach to further subdivide systems into more specific classes according to criteria specific to this particular field, this classification scheme affords interdisciplinary search, which is generally left out of generic indexing systems. This paper presents the resulting classification scheme and showcases its usefulness for classifying and tracking the evolution of research.

Introduction

The field of Case-Based Reasoning in the Health Sciences (CBR-HS) has seen a tremendous growth in the last decade. Six special conference workshops have been held consecutively beginning in 2003 focused solely on this topic and are accessible through the Cbr-biomed.org Web portal (Bichindaritz and Reed 2007). A special issue on CBR-HS was published in the Journal of Artificial Intelligence in Medicine (Bichindaritz 2006) and a second one in the Computational Intelligence journal (Bichindaritz and Marling 2006). Most recently, yet another special issue has been published in the Applied Intelligence journal (Bichindaritz et al. 2008). Moreover CBR-HS papers are often published in different artificial intelligence and health informatics journals and conferences because they are interdisciplinary. With the increase in amount of information, there is a need to index documents in such a manner as to facilitate search and track evolution. One solution is to create a classification scheme.

There are many benefits to any computer based information classification scheme. The most obvious is the ability to quickly find what one is looking for, and with precision not allowed by broad indexing schemes – such as for CBR-HS the field of case-based reasoning, health informatics, and/or both. As the amount of information grows at a fast pace, the ability to quickly find relevant information becomes more desirable. Another benefit is the ability to compare one body of information to another by similarity. In a very simple keyword classification scheme, the number of common keywords could be counted to determine similarity between two articles. However, this approach is not successful if the keywords provided are too broad, as is often the case. In particular, a good classification scheme permits documents to be grouped into finer-grained or coarser-grained categories, such as for example "psoriasis" or "skin disease", and to track the evolution of research along different categories over time. As a matter of fact, associating keywords to documents serves as a preprocessing step for text mining endeavors.

No classification scheme is known to exist that provides all of these benefits to the field of case-based reasoning (CBR) in the Health Sciences. Recently Greene et al. could not identify a CBR-HS group by automatic clustering methods (Greene et al. 2008). As a result, this paper attempts to develop a new scheme, called the CBR-HS classification scheme, based on a thorough analysis of the current state of research in the field. This paper details the methodology followed, provides examples of using this system, and attempts to classify recent literature in CBR-HS to identify the benefits of the classification.

Background

The specific application of CBR to the health sciences has been discussed in several surveys (Gierl et al. 1998, Schmidt et al. 2001, Nilsson and Sollenborg 2004, Bichindaritz 2006, Holt et al. 2006).

The history of classifying information is also long and rich. Entire volumes have been published on the subject. This section will focus on the two main areas of CBR-HS, computer science and health sciences. One of the oldest and most established methods for classifying information is the Dewey Decimal System (Dewey 1976). Its novel use of numbers to represent categories and dots to separate subcategories has been emulated in several other schemes. However, the use of numbers was mainly for compactness, which is not as valuable in modern classification systems.

The Association for Computing Machinery (ACM) Computing Classification System (CCS) is one of the oldest schemes in the computer science field and has undergone several major revisions, mostly in the underlying content and less in the actual framework (Coulter et al. 1998, ACM 2007). Like the Dewey Decimal System, it links categories in a topology with descriptors coded using numbers and dots for the branches in the tree. However, the final leaf of the classification code is generally uncoded and more than one code can be used, so they are not unique. The ACM CCS code *D.2.5 Tracing* would be deciphered as follows:

```
(D) Software
D.(2) Software Engineering
D.2.(5) Testing and Debugging
D.2.5 (Tracing)
```

One of the most frequently used classification schemes in the health sciences is the Medical Subject Headings (MeSH). It was created in 1960 by the U.S. National Library of Medicine (Lipscomb 2000). Like the ACM CCS, it uses a topology where broader categories are narrowed down with each branch and branches are represented by dots. However, the leaves are not uncoded in this scheme. As an example, the MeSH code *C10.228.140.380.100* would be deciphered as follows:

```
(C10) Nervous System Diseases
C10.(228) Central Nervous System Diseases
C10.228.(140) Brain Diseases
C10.228.140.(380) Dementia
C10.228.140.380.(100) Alzheimer's Disease
```

Very interesting work has been accomplished with MeSH. One such project involves the calculation of the conceptual distance between articles using MeSH terms (Ontrup et al. 2003). Though this is one of the desired benefits of CBR-HS classification scheme, these systems can probably only be mapped to a few MeSH terms at most. Additionally, in support of the idea that classification schemes benefit searching, a recent study showed searching based on MeSH terms was much more efficient than text word searching alone (Chang et al. 2006).

Relating computer science, the Information Systems Research Literature (ISRL) categories were developed in the late 1980s to create an agreed upon list of keywords describing the field, introduce a common language, prevent the proliferation of synonyms, and enable the development of better research databases (Barki et al. 1993).

The main limitation of the aforementioned schemes is they characterize information only along one dimension, that being the topic (Ramesh et al. 2004). Even with established schemes, many journals and organizations still use a rudimentary keyword classification scheme where the author merely lists keywords associated with the article. More recently, a proposed unified classification for the computing disciplines (Vessey et al. 2005) took a step in the right direction by looking at more than one dimension of an article. However, the extra dimensions deal mostly with the research process and it is unclear how beneficial that is, as some CBR-HS systems may be more application than research orientated.

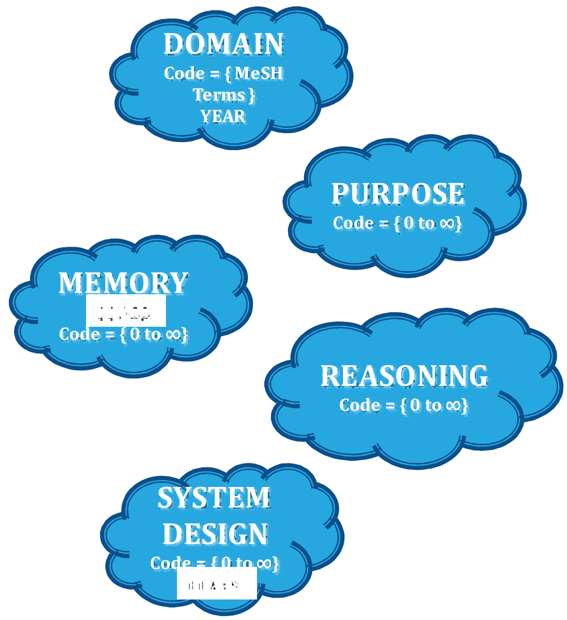

Fig. 1. CBR Health Sciences tiered classification scheme

Classification System

CBR-HS is an interdisciplinary field, which has motivated studying both classification schemes in computer science and in the health sciences. The CBR-HS scheme diverges from most classification schemes in that a tree structure is not used for the primary means of classification. Instead, a tiered approach is taken with each tier representing a category that contains a separate set of descriptors. From the top down each set of descriptors should decrease in size. Each descriptor, within a set, is represented by a unique code and follows a tree structure. Each code should be made up of an, theoretically, infinite number of digits, or in the top level case letters. In cases where specific aspects of a system are in finite number, flags can be used instead of codes to represent the descriptors.

Other than for compactness, use of codes has other benefits. If numbers are used in a meaningful way, then mathematical operations can be performed on them. To obtain the greatest benefit, numbers should be arranged so that similar topics are grouped more closely together. In fact, it would be encouraged to add additional numbers if one topic overlaps another. For instance, 11 would be more related to 12 than to 20. However, since the codes represent concepts, their semantic content could yield to alternate ways of connecting them.

A general depiction of the CBR-HS classification scheme is presented in Figure 1. As can be seen, there are five distinct categories (domain, purpose, memory and case management, reasoning, and system design), which are defined in the following sections, and each category contains at least one set of codified descriptors.

Domain

The range of domains in the health sciences fields is vast and, as a result, it was chosen as the first level of classification. However, rather than creating a new set of descriptors, it is proposed to use the MeSH descriptors, of which there are over 24,000 that cover just about every aspect of the health sciences. The MeSH codes (see Table 1) can be used unmodified and separated by commas if more than one is appropriate.

Table 1. Sample Domain Classifications

Code	Domain
C04	Cancer
C14.907.489	Hypertension
F04.096.544	Psychiatry
A11.872	Stem Cells
C10.228.140.380.100	Alzheimer’s
B05	Fungi

Choosing the largest field first allows for the most discrimination at the beginning. That is, the first field is able to discriminate more papers from one another than any other field. However, as the ordering of the tiers does not matter, this methodology only benefits systems that order items based on a left to right reading.

Along with the domain, another primary means of discriminating the relevance of an article is its publication date. More recent articles often have more relevance, as the understanding of topics develops over time. Since the date plays no real role in classifying an article, the date has no field of its own, but instead is combined with the Domain.

Purpose

The purposes, or tasks, of CBR systems have been thoroughly discussed in many articles summarizing the CBR-HS domain. One of the first papers to survey the field in 1998, by Gierl et al., used the purpose as the primary means to subdivide the different systems (Gierl et al. 1998). Gierl et al. specified four main purposes: diagnosis, classification, planning, and tutoring. Later, Holt et al. (2006) and Nilsson and Sollenborn (2004) used the same four descriptors. In the early years the majority of systems were diagnostic in nature, but in recent years more therapeutic and treatment systems have been developed (Schmidt 2007). Table 2 presents examples of purpose classifications. Planning has been replaced here by treatment since most of the time planning refers to treatment planning. However, planning tasks may involve not only treatment but also other aspects such as diagnosis assessment, which often consists of a series of exams and labs orchestrated in a plan. Planning is a classical major task performed by artificial intelligence systems. Therefore planning is listed in our system as a design option to add to the treatment choice in the purpose dimension.

Table 2. Sample Purpose Classifications

Code	Purpose
10	Medical Purpose
10.1	Decision Support
10.1.1	Diagnosis
10.1.2	Treatment
10.1.3	Prognosis
10.1.4	Follow-up
10.1.5	Classification
10.2	Tutoring
10.3	Epidemiology
10.4	Research support
20	Research Purpose
20.1	Formalization
20.2	Evaluation
20.2.1	System Level Testing
20.2.2	Pilot Testing
20.2.3	Clinical Trial
20.2.4	Routine Clinical Use
20.3	Concept
20.4	Method
20.5	Survey
30	Bioinformatics Purpose
30.1	Proteomics
30.2	Phylogenetics
30.3	Genomics
40	Research Theme

CBR systems generally support either medical clinical work or research, or bioinformatics research. Therefore we have added these as top level purpose categories. In the clinic, decision support systems support mostly diagnosis, treatment, prognosis, follow-up, and/or classification, such as in image interpretation. Well known diagnostic systems include CARE-PARTNER (Bichindaritz et al. 1998), and AUGUSTE (Marling and Whitehouse 2001). Well known classification systems include PROTOS (Bareiss et al. 1987) and IMAGECREEK (Grimnes and Aaomdt 1996). Well known treatment planning systems include T-IDDM (Montani et al. 2000). Several systems provide multi-expertise, such as CARE-PARTNER (Bichindaritz et al.

1998) affording diagnosis and treatment planning, which can be represented by listing the different purposes. Well known tutoring systems include ICONS (Gierl 1993).

More recent articles differentiate between the purpose of the system developed, which is generally a clinical purpose, from the purpose of the research paper, which can be, among others, a survey paper or a classification paper like this one. Some papers focus on formalization such as KASIMIR (Lieber et al. 2008). Among these, the evaluation of a system can be performed more or less thoroughly. This is an important dimension to note about a research paper: whether a system was tested only at the system level, which is the most frequent, at the pilot testing level, at the clinical trial level, or finally whether the system is in routine clinical use.

Finally, a paper is generally identified by a research theme by its authors. By indexing a set of 326 papers currently in our database, we have identified major research themes, such as CBR and electronic medical records (EMR), knowledge morphing, CBR and clinical guidelines, or application of a generic CBR framework.

Table 3. Sample Memory and Case Management Classifications

Code	Memory Organization
10	Flat
20	Hierarchical
20.1	Decision Tree
20.2	Concept Lattice
20.3	Conceptual Clustering Tree
Case Representation Flag	
I	Images
S	Signals
T	Time Series
A	Text
M	Microarray
V	Attribute/Values
W	Mass Spectrometry
Memory Structures Flag	
G	Ground Cases
P	Prototypical Cases
L	Clusters
O	Concepts

Memory and Case Management

This is a very broad category and could easily be subdivided. It encompasses both how the cases are represented and how they are organized in memory for retrieval purposes and more. As a result, it is made up of more than one code. The first part of the code represents the format of the cases. The primary types being images, signals, mass spectrometry, microarray, time series data and regular attribute/values pairs, which is used by the majority of the systems. Since this set is finite, a flag can be used and to keep the flag separate from the code a letter should be used, as opposed to a number for the code. If there is a differentiation between upper and lower case letters, this means the flag can contain at most 52 elements. Similar to the different formats of data are the flags that represent what kinds of memory structures the CBR system uses to represent the data, such as ground cases (G), prototypical cases (P), clusters (L), or concepts (O). Lastly, when it comes to memory management there are potentially an infinite number of possibilities. The main types, however, represent how the memory is organized, whether it is flat or hierarchical, what kind of hierarchical structure, such as decision tree, concept lattice, conceptual clustering tree, or others. A sample of memory management codes and flags can be seen in Table 3. More than one code can be used in the CBR-HS scheme by simply separating them in the code with a comma.

Table 4. Sample Reasoning Classifications

Code	Reasoning
10	Retrieve
20	Reuse
20.1	Adaptation
20.2	Interpretation
30	Revise
40	Retain

Reasoning

This category groups the inferential aspects of the CBR. Classically, retrieve, reuse (and its variations known as adaptation and interpretation), revise, and retain have been described. Nevertheless, researchers have often added many more aspects to the inferences, such that it is best to keep this category open to important variations (see Table 4). Each of these parts of the reasoning cycle can be hierarchically refined.

System Design

The construction of the CBR system specifies what technologies it uses. This area of classification may not seem intuitive at first, but upon the examination of CBR systems it can be seen that many use a combination of technologies, not just case-based reasoning. The most common technology used in conjunction with CBR is rule-based reasoning; however some systems combine CBR with information retrieval, data mining, or other artificial intelligence methods. See Table 5 for an example of different possible construction classifications. If the construction of the system does use additional technologies, a flag should be appended to the end of the code to denote whether the case-based reasoning is executed separately, the flag being a T for true or F for false. Also, an additional flag is used to designate CBR's role(s) in the system, as it may be possible that CBR plays only a small part in the overall system, corresponding to an S flag. Here also several designs can be added by separating them with a comma.

It may be possible to use the ACM CCS code in place of the custom ones proposed here. However, that scheme is currently too vague, as most of the technologies presented here would simply be classified under Artificial Intelligence (I.2).

Table 5. System Design Classifications

Code	Construction
10	Pure CBR
20	Rule Based Combination
30	Model Based Combination
40	Data Mining Combination
40.1	Conceptual Clustering
40.2	Neural Networks
40.3	Nearest Neighbor
40.4	Decision Tree
40.5	Bayesian Networks
50	Planning Combination
60	Information Retrieval Combination
70	Explanation Combination
CBR Role Flag	
P	Primary Technology
S	Secondary Technology
E	Equivalent Role Technology
CBR Additional Technology Flag	
T	CBR is Separate
F	CBR is Combined

Example

To demonstrate the CBR-HS classification scheme, an example will be presented using the CARE-PARTNER (Bichindaritz et al. 1998) system. The first step is to determine the domain. Two good candidates would be cancer and stem cells. Either one or both can be used and, in theory, the more the better. Looking up the terms in MeSH gives *C04* for the former and *A11.872* for the latter. The publication year is 1998, so the entire domain portion of the CBR-HS code would be C04,*A11.872#1998*. The purpose of the system is diagnosis, treatment, and follow-up, so this give a category code of *10.1.1, 10.1.2, 10.1.4*. Next, the memory and case management must be analyzed. The flags are fairly easy to set. The case format is attribute/value and text, it uses prototypes to some extent, as well as ground cases. This gives an overall flag of *V,AP,G*. The case management and retrieval uses a hierarchical conceptual clustering tree, which gives a code of *20.3*. The system focuses on retrieve, reuse, and retain, which gives a code of *10, 20, 40*. Finally, the system is CBR combined with conceptual clustering, but CBR is the primary methodology, so the system design code contains *P*, in addition to *40.1* for conceptual clustering and *F* for CBR is combined.

However all these characteristics of the system do not encompass the research theme(s), which here can be identified as *CBR and clinical guidelines*, or *40.2*. It is added to the purpose. This example illustrates that in many cases system characteristics and research theme are different concepts. If the clinical purpose or medical research purpose are important to track, research themes do not automatically emerge from these characteristics.

Putting everything together gives a CBR-HS code of *C04,A11.872#1998:10.1.1,10.1.2,10.1.4,40.2:V,AP,G20.3: 10,20,40:40.1PF*. Of course users have access to the concepts represented by this code, which yields a much more readable information.

Creating a code for an article is not as complicated as it might seem. Given a proper interface and knowledge of the system, a code could be generated in a matter of minutes.

Discussion

One of the key concerns of any classification system is its stability. If it is not maintained over time it can easily fail. Keeping this in mind, it would be advisable to create a committee to oversee the schemes construction and maintenance. Just the simple idea of classifying information by more than one dimension is a very powerful concept that should be considered when creating any new classification scheme, particularly in multidisciplinary domains. Three of the five categories presented here can easily be extended to other domains. The domain, or topic, is how most information is already classified. The purpose, or goal, should be applicable to any research. Lastly, the method of construction or system design is present in many fields, but not as universal as the domain or purpose.

One main advantage of this CBR-HS classification is that it affords for expansion by adding subcategories and reorganization by merging of categories. This is particularly important for the Reasoning tier. In addition, in this system, both theoretical papers and applied papers can be coded, since not all tiers need to be documented. For example, a theoretical paper could ignore the System Design tier, and focus more on the Purpose tier.

A central repository, with a web interface, would greatly facilitate the searching and sharing of information in the CBR-HS domain on a Website. Additionally, tracking the evolution of CBR-HS systems becomes a simple task with all of the codes in one repository. Given the date information in the codes, statistics, such as what percentage of systems use prototypes, can be compiled over time.

Although the hierarchical organization of this classification is classical, it represents a solid start at tracking the CBR-HS research literature, and could be later transformed into a graph-like indexing structure, which would open more refined search possibilities.

The power of this classification theme for searching and trend tracking has been evaluated on a subset of papers, providing the tables in this article. Detail of these trends and statistics will be presented in another paper.

Conclusion

The CBR-HS classification system is being incrementally built. The different categories and each category's list of descriptors is by no means exhaustive. However, much understanding of the domain has already been gained in terms of medical domains, research themes, and CBR characteristics. The purpose of this paper is not to impose a classification system on the researchers, but instead present ideas open to discussion. When developing a classification scheme there are many tradeoffs. The more complex the system, the better an automated system will be able to analyze it, and for example to identify the research themes like Greene et al. have done for CBR in general (Greene et al. 2008). At the same time the more difficult to use. Adding extra dimensions is the key contributor to the complexity of the scheme, but the benefits gained far outweigh the additional effort needed to catalog CBR-HS information. Over time a robust system can be developed that will benefit everyone in the community. If proven useful, this system could be extended to other interdisciplinary domains such as data mining in the health sciences, for example. The CBR aspects could also be expanded to encompass all CBR papers indexing.

References

ACM. The ACM Computing Classification sSystem. 1988. http://www.acm.org/class/1998/ (1998) Date of publication: 1998.

Bareiss, E.R.; Porter, B.W.; and Wier, C.C. 1987. Protos: an exemplar-based learning apprentice. In *Proceedings of the Fourth International Workshop on Machine Learning*, 12-23. Los Altos, CA: Morgan Kaufmann.

Barki, H.; Rivard, S.; and Talbot, J. 1993. A Keyword Classification Scheme for IS Research Literature: An Update. *MIS Quarterly* 17(2): 209–226.

Bichindaritz, I.; Kansu E.; and Sullivan, K.M. 1998 Case-based reasoning in care-partner: gathering evidence for evidence-based medical practice. In Smyth, B. and Cunningham, P. eds *Proceedings of the 4th European Workshop on CBR*, 334–345. Berlin, Germany: Springer-Verlag.

Bichindaritz, I.; Marling, C. 2006. Case-based reasoning in the health sciences: What's next? *Artificial Intelligence in Medicine* 36(2): 127–135.

Bichindaritz, I., and Marling, C. 2006. Introduction to the special issue on case-based reasoning in the health sciences. *Computational Intelligence* 22(3–4): 143–147.

Bichindaritz, I.; Montani, S.; and Portinale, L. 2008. Special issue on case-based reasoning in the health sciences, *Applied Intelligence* Vol. 28, Issue 3: 207-209.

Bichindaritz, I., and Reed, J. 2008. Cbr in the health sciences workshops. http://www.cbrbiomed.org/workshops/ Date of publication: October 1, 2007. Date retrieved: February 1, 2008.

Bichindaritz, I. 2006. Case-based reasoning in the health sciences. *Artificial Intelligence in Medicine* 36(2): 121–125.

Chang, A.A.; Heskett, K.M.; and Davidson, T.M. 2006. Searching the literature using medical subject headings versus text word with pubmed. *The Laryngoscope* 116(2): 336–340.

Coulter, N.; French, J.; Glinert, E.; Horton, T.; Mead, N.; Rada, R.; Ralston, A.; Rodkin, C.; Rous, B.; Tucker, A.; Wegner, P.; Weiss, E.; and Wierzbicki, C. 1998. Computing classification system 1998: Current status and future maintenance report of the ccs update. *Computing Reviews* January 1998.

Dewey, M. 1876. *A Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a Library.* KingsPort, Tennessee: .Kingsport Press Inc.

Gierl, L.; Bull, M.; and Schmidt, R. 1998. CBR in Medicine. In *Case-based reasoning technology. From foundations to applications*, 273–297. Berlin, Germany: Springer-Verlag.

Gierl, L. 1993. Icons: Cognitive Basic Functions in a Case-Based Consultation System for Intensive Care. In Andreassen, S et al. eds *Proceedings of the 4th Conference on Artificial Intelligence in Medicine Europe*, 230–236. Amsterdam, Neth: Elsevier Science Publishers B.V.

Greene, D.; Freyne, J.; Smyth, B.; and Cunningham, P. An Analysis of Research Themes in the CBR Conference Literature. In *Proceedings of the 9th European Conference, ECCBR 2008*, 18-43. Berlin: Springer-Verlag.

Grimnes, M.; Aamodt, A. 1996. A two layer case-based reasoning architecture for medical image understanding. In Smith, I and Faltings, B eds *Proceedings of the Advances in Case-Based Reasoning. Third European Workshop, EWCBR-96,* 164–178. Berlin, Germany: Springer-Verlag.

Holt, A.; Bichindaritz, I.; Schmidt, R.; and Perner, P. 2006. Medical applications in case-based reasoning. *The Knowledge Engineering Review* 20(3): 289–292.

Lieber, J.; dAquin, M.; Badra, F.; and Napoli, A. 2008. Modeling adaptation of breast cancer treatment decision protocols in the kasimir project. *Applied Intelligence* Vol. 28 Issue 3: 261-274.

Marling, C.R., and Whitehouse, P.J. 2001. Case-based reasoning in the care of Alzheimer's disease patients. In *Proceedings of the International Conference of Case-Base Reasoning, ICCBR-01* 702–715. Berlin, Germany: Springer-Verlag.

Montani, S.; Bellazzi, R.; Portinale, L.; and Stefanelli, M. 2000. A multi-modal reasoning methodology for managing iddm patients. *International Journal of Medical Informatics* 58–59: 243–256.

Nilsson, M., and Sollenborn, M. 2004. Advancements and trends in medical case-based reasoning: An overview of systems and system development. In *Proceedings of the 17th International FLAIRS Conference,* 178–183. Menlo Park, Calif.: AAAI Press.

Ontrup, J.; Nattkemper, T.W.; Gerstung, O.; and Ritter, H. 2003. A mesh term based distance measure for document retrieval and labeling assistance. In *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 1303–1306. Piscataway, NJ: IEEE Press.

Ramesh, V.; Glass, R.L.; and Vessey, I. 2004. Research in computer science: an empirical study. *Journal of Systems and Software* 70(1–2): 165–176.

Schmidt, R.; Montani, S.; Bellazzi, R.; Portinale, L.; and Gierl, L. 2001. Case-based reasoning for medical knowledge-based systems. *International Journal of Medical Informatics* 64(2-3): 355–367.

Schmidt, R. 2007. Case-Based Reasoning in Medicine Especially an Obituary on Lothar Gierl. In *Advanced Computational Intelligence Paradigms in Healthcare*. 63–87. Berlin Heidelberg New York: Springer-Verlag.

Vessey, I.; Ramesh, V.; and Glass, R.L. 2005. A unified classification system for research in the computing disciplines. *Information and Software Technology* 47(4): 245–255.