

A Hierarchical Approach to the Automatic Categorization of Medical Documents

Luciano R. S. de Lima^{1,2} Alberto H. F. Laender¹ * Berthier A. Ribeiro-Neto¹ †

¹Computer Science Department
Federal University of Minas Gerais
31270-901 - Belo Horizonte - MG - Brazil
{luciano, laender, berthier}@dcc.ufmg.br

²Medical Informatics Group
Sarah Hospital Network
30510-000 - Belo Horizonte - MG - Brazil
luciano@bhz.sarah.br

Abstract

A common problem in many medical organizations is the categorization of medical documents, such as inpatient discharge summaries, with the aim of facilitating their archiving and later retrieval. For this purpose, the medical organizations usually adopt coding standards proposed by international health organizations. In this paper, we first describe a hierarchical model we proposed for the automatic categorization of medical documents. Our model is based on the International Code of Diseases (ICD) proposed by the World Health Organization and takes advantage of the hierarchical structure of the ICD to categorize the documents with high precision. We then present two algorithms for categorization of medical documents (i.e., the actual code assignment), one based on the classic vector space model and the other based on our hierarchical model, and discuss the results of a set of experiments we carried out to compare their retrieval performance. Our results show that the algorithm based on our hierarchical model outperforms the algorithm based on the classic vector space model and that it presents quite good precision and recall figures for the problem of code assignment to medical documents.

1 Introduction

A common problem in many medical organizations is the categorization of medical documents, such as inpatient discharge summaries, with the aim of facilitating their archiving and later retrieval. For this purpose, the medical organizations usually adopt coding standards proposed by international health organizations. In general, the categorization is done by comparing terms extracted from the documents with the terms that compose the vocabulary of the adopted coding standard in order to determine the appropriate codes that might be assigned to them. This procedure is usually

*Partially supported by CNPq under individual research grant number 300959/85-0 and by FINEP/PRONEX under Project SIAM grant number 76.97.1016.00.

†Partially supported by CNPq under individual research grant number 300188/95-1 and by FINEP/PRONEX under Project SIAM grant number 76.97.1016.00.

carried out manually by coding specialists, and constitutes a tedious and very costly task which requires attention and careful interpretation, besides being error prone.

In this paper, we first describe a hierarchical model we proposed [5] for the automatic categorization of medical documents. Our model is based on the International Code of Diseases (ICD) [6] proposed by the World Health Organization and takes advantage of the hierarchical structure of the ICD to categorize the documents with high precision. Despite being based on the ICD, our model is quite general and, we believe, could be equally applied to other hierarchical coding standards such as the Systematized Nomenclature of Medicine (SNOMED), the International Classification of Primary Care (ICPC), and the Read Clinical Codes (RCC) [1]. We then present two algorithms for automatic code assignment, one based on the classic vector space model [8] and the other based on our hierarchical model, and discuss the results of a set of experiments we carried out to compare their retrieval performance. Our experiments are based on a medical database from a Brazilian hospital, and therefore we use a Portuguese version of the International Code of Diseases, version 9 (ICD-9) [6]. Our results show that the algorithm based on our hierarchical model outperforms the algorithm based on the classic vector space model and that it presents quite good precision and recall figures for the problem of code assignment to medical documents.

Other approaches have been proposed in the literature to address the problem of code assignment to medical documents but they differ from our work substantially. In [3, 4], the problem is treated as a classification problem which is solved by combining three classifiers based on probabilistic models [7] for the code assignment task. In [9], the problem is addressed through the usage of natural language processing techniques. Both approaches present good precision and recall figures in some situations but have some disadvantages when compared to ours. In the first case, good results are dependent on a large number of medical documents which have been previously classified and can be used as a training set. In the second case, an excessively complex approach is adopted to address a problem whose domain vocabulary is clearly small when compared to the vocabulary of any existing language. In addition, contrary to our work, in both approaches the hierarchical structure of the coding standard and the knowledge of the coding specialists are completely ignored.

The remainder of the paper is organized as follows. Section 2 presents an overview of the International Code of Diseases, version 9, which is the coding standard adopted in our experiments. Section 3 describes our hierarchical model

for automatic code assignment. Section 4 describes the two code assignment algorithms we implemented in our experiments. Section 5 describes the set of experiments we carried out to compare the retrieval performance of the two code assignment algorithms and analyzes their results. Finally, Section 6 concludes the paper.

2 ICD: International Code of Diseases

In this section, we briefly describe the structure of the *international code of diseases*, which we refer to as ICD. This structure reflects the hierarchical nature of the ICD coding scheme which is important here because, distinctly from previous approaches for automatic code assignment [3, 4, 9], it heavily influences the code assignment algorithm we propose.

The ICD code scheme, which has been proposed by the World Health Organization (WHO), defines a vocabulary of medical terms for describing diseases, injuries, and death causes. This vocabulary is largely used by health organizations throughout the world for assigning ICD codes to a large variety of medical documents. The ninth version of the ICD, called ICD-9, is the one most widely used throughout the world and is adopted in our study (as in [3, 4]). Since our experiments are based on a medical database from a Brazilian hospital, we use a Portuguese version of the ICD-9 [6].

The ICD-9 is organized hierarchically in two documents: a tabular list of codes and an alphabetical index of vocabulary terms. The first ICD-9 document, the *tabular list of codes*, is divided in four parts called section, grouping, category, and subcategory. Figure 1 illustrates a small portion of the ICD-9 tabular list of codes. The level *I* is a section, the level *I.1* is a grouping, the level 001 corresponds to the category *cholera*, and the level 001.1 corresponds to the subcategory *cholera due to the cholerae vibrio el Tor*. Despite these four parts, in general, only the levels of category and subcategory are used as reference codes. The other two levels are too generic and do not provide proper code descriptions. Thus, in this study we focus on the automatic assignment (to inpatient discharge summaries) of ICD-9 category and subcategory codes only (as in [3, 4]).

<i>I</i>	<i>Infectious and Parasitic Diseases</i>
<i>I.1</i>	<i>Intestinal Infectious Diseases</i>
<i>001</i>	<i>Cholera</i>
<i>001.0</i>	<i>due to Vibrio cholerae</i>
<i>001.1</i>	<i>due to Vibrio cholerae el tor</i>
<i>001.9</i>	<i>unspecified</i>

Figure 1: ICD-9 tabular list of codes (translated to English).

Complete descriptions of category and subcategory codes are found in the second ICD-9 document, the *alphabetical index* of vocabulary terms, as shown in Figure 2. An entry point in this index is marked by a term which is not preceded by a dash mark. For instance, in Figure 2, *cholera* is an entry to which a classification code might be assigned to. In this case, the code 001.9 is assigned to the entry point

cholera. Dash marks are used to indicate terms which are hierarchically dependent of an entry. For instance, *antimomial* is a hierarchical descend of *cholera* in Figure 2. To the combination *cholera antimomial* is assigned the code 985.4 which is more specific than the code 001.9. For simplicity, we refer to a combination of hierarchically depend terms which starts at an entry point as a *codepath*. In Figure 2, the code 001.0 is associated to the codepath *classic cholera* while the code 001.1 is associated to the codepaths *cholera el Tor* and *cholera vibrio cholerae el Tor*. Notice that a same code might be associated to two or more codepaths.

Furthermore, notice that the alphabetical index in Figure 2 includes a subcategory code 001.0 relative to *classic cholera* which is not present in the tabular list of codes. Also, the entry point for the category *cholera* includes not only the code category 001 but also the subcategory 001.9 which corresponds to *cholera due to a non-specified (or unknown) cause*. This subcategory code does not necessarily appear explicitly in the tabular list of codes.

<i>Cholera 001.9</i>
<i>-Antimomial 985.4</i>
<i>-Classic 001.0</i>
<i>-el Tor 001.1</i>
<i>-Vibrio</i>
<i>--Cholerae 001.0</i>
<i>... el Tor 001.1</i>

Figure 2: ICD-9 alphabetical index (translated to English).

Thus, since the ICD-9 alphabetical index is more complete, our automatic code assignment algorithm is based on it.

3 A Hierarchical Model for Automatic Code Assignment

In this section, we describe a hierarchical model we proposed [5] for automatic code assignment. This model takes advantage of the hierarchical topology of the code structure which provides for high precision in the code assignment task as demonstrated in Section 5. In this study we focus on ICD-9 codes. However, our model is equally applicable to other hierarchical coding schemes such as the Systematized Nomenclature of Medicine (SNOMED), the International Classification of Primary Care (ICPC), and the Read Clinical Codes (RCC) (see [1] for details). Since medical classification schemes tend naturally to be hierarchical, our model is quite general for coding of medical documents.

The model includes several components as follows.

Definition 1 A linguistic vocabulary V is a set $V = \{t_1, t_2, \dots, t_{nv}\}$, $nv \geq 1$, which defines all terms t_i of the universe of discourse of an specialized area of study. Usually, $V = V_s \cup V_g$, where V_s is a vocabulary of specialized terms and V_g is a vocabulary of generic (or non-specialized) terms.

Definition 2 A set C of classification codes is given by $C = \{c_1, c_2, \dots, c_{nc}\}$, $nc \geq 1$, where each c_i is a classification code.

A *hierarchical coding structure CS* is a direct acyclic graph (DAG) in which the following conditions are verified: (a) to each node of this DAG is associated a term of the specialized vocabulary V_s and (b) to each edge of this graph is associated a set of synonyms and a set of classification codes. Synonyms and classification codes are associated to edges (instead of nodes) because their utilization in the code assignment process is dependent on the terms associated to the pair of nodes for that edge. Definition 3 formalizes this concept.

Definition 3 A *hierarchical coding structure CS* is a direct acyclic graph given by $CS = (N, E)$ where $N = \{n_1, n_2, \dots\}$ is a set of nodes (i.e., vertices) and E is a set of directed edges connecting nodes of N . We use the notation $e_{ij} = (n_i, n_j)$ to refer to the edge e_{ij} pointing from the node n_i to the node n_j . The nodes n_i and n_j are said to be the endpoints of e_{ij} . The node n_i is said to be a parent node of n_j . Nodes without parents are called root nodes. For convenience, we define a node n_0 which is used as a (false) parent of all root nodes. To each node n_i , $n_i \in N$, is associated a single term from the specialized vocabulary V_s . Let $t(n_i)$ be the term associated to the node n_i . To each edge e_{ij} are associated two sets: a set S_{ij} of synonyms and a set $C_{ij} = \{c_1, c_2, \dots\}$ of classification codes, where $c_i \in C$. The set S_{ij} specifies all synonyms of the term $t(n_j)$ in the context of the term association $(t(n_i), t(n_j))$. The set C_{ij} specifies a set of classification codes associated to the term $t(n_j)$ in the context of the term association $(t(n_i), t(n_j))$.

Notice that synonyms and classification codes for a node n_j depend on the parent node which is traversed to reach n_j . That is the reason for associating synonym and classification code sets to the edges in the graph. Notice also that synonyms and classification codes associated to a root node n_k are given by the sets S_{0k} and C_{0k} . Further, if neither synonyms nor classification codes are associated to an edge e_{ij} then $S_{ij} = \{\}$ and $C_{ij} = \{\}$.

Definition 4 A *path P* in the hierarchical coding structure CS is a sequence of edges $e_{k_1 k_2}, e_{k_2 k_3}, \dots, e_{k_{p-1} k_p}$ in which (a) the first node n_{k_1} in the sequence is a root node and (b) consecutive edges share a common endpoint (as indicated by the notation we adopt). The number p of nodes in the path P is its path length. Further, the edge $e_{k_{p-1} k_p}$ is called the terminal edge of P .

Definition 5 An *acronym dictionary* is a set $A = \{a_1, a_2, \dots, a_{n_a}\}$, $n_a \geq 0$. Each element a_i is a pair given by $(acron_i, str_i)$ where $acron_i$ is an acronym of the linguistic vocabulary V and str_i is a string (composed solely by terms of the specialized vocabulary V_s) which can be referred to by $acron_i$.

Acronyms are useful because doctors use them frequently when writing down medical documents.

Definition 6 A *synonym dictionary* is a set $S = \{s_1, s_2, \dots, s_{n_s}\}$, $n_s \geq 0$. Each element s_i is a pair given by (sin_i, str_i) where sin_i and str_i are synonym strings. Further, the string sin_i is composed of terms of the linguistic vocabulary V while the string str_i is composed of terms of the specialized vocabulary V_s .

As acronyms, synonyms are useful because they are popular among doctors.

A query in our model is simply a medical document to which a classification code has to be assigned. This medical document is usually composed of sections as follows.

Definition 7 A *specialized query* is a medical document represented by a set $Q = \{q_1, q_2, \dots, q_{n_q}\}$, $n_q \geq 1$. Each element q_i is a section of the medical document Q and is represented by a tuple (l_i, str_i) where l_i is a section label and str_i is a string composed of terms of the linguistic vocabulary V .

Notice that, in our model, the ordering among the sections of a medical document is not important.

Definition 8 A *classification code assignment to a specialized query Q* is a set of tuples $R_i = (Q, c_i, r_i)$ where $c_i \in C$ is a classification code assigned to the query Q and r_i is a rank which quantifies a degree of confidence in this classification. In our model, $0 \leq r_i \leq 1$.

Thus, a code assignment is simply the association of a sorted list of classification codes to a medical document (which is treated as a query). In Section 4 we discuss how to apply the above specialized hierarchical model to the ICD-9 code assignment problem.

4 Automatic Code Assignment Algorithms

In this section, we describe two algorithms for the automatic assignment of ICD-9 codes to medical documents. The first one is based on the classic vector space model [8] and not considers the hierarchical structure of the ICD-9 alphabetical index. The second one takes into account the hierarchical topology of the ICD-9 index and is based in our hierarchical model described in Section 3.

4.1 Code Assignment based on the Vector Space Model

The vector space model considers that documents and queries are indexed by keywords. To each keyword k_i in a document d is assigned a weight w_{id} which is usually based on a tf-idf (i.e., term frequency and inverse document frequency) scheme [8]. To each keyword k_i in a query q is also assigned a weight w_{iq} . The similarity (or rank) $sim(d, q)$ of the document d with respect to the query q can be computed, for instance, by the cosine of the angle between the two vectors as given by Equation 1.

$$sim(d, q) = \frac{\sum_{\forall i} w_{id} \times w_{iq}}{\sqrt{\sum_{\forall i} w_{id}^2} \times \sqrt{\sum_{\forall i} w_{iq}^2}} \quad (1)$$

We can apply the vector space model to the ICD-9 code assignment problem as follows:

1. Each medical document is interpreted as a query Q while the ICD-9 codes are viewed as documents to be retrieved. Thus, our document collection is the set of all ICD-9 codes.
2. To each medical document is assigned a vector of terms extracted from its text. Each acronym is replaced by its respective term string. All stop words are filtered out.
3. To each classification code c_i in the ICD-9 alphabetical index is associated a set of vectors of terms. Each of these vectors is composed by the terms in a codepath (see Section 2) for the code c_i . Again, all stop words are filtered out.

4. The weights are computed as *tf-idf* (i.e., within-document frequency and inverse document frequency) factors in standard fashion [8].

Since more than one vector of term weights might be assigned to a same classification code c_i (because, as discussed in Section 2, the ICD-9 alphabetical index might associate more than one codepath to c_i), we compute the similarity $sim(c_i, Q)$ between the code c_i and the medical discharge summary Q as the maximum of all similarities between Q and each of the term weight vectors associated to c_i .

4.2 Code Assignment based on the Hierarchical Model

The hierarchical model can be used to model the assignment of ICD-9 codes as follows:

1. As for the vector model, each medical document is interpreted as a query Q while the ICD-9 codes are viewed as documents to be retrieved.
2. The ICD-9 hierarchical alphabetical index is modeled as a hierarchical coding structure which we refer to as $CS-9=(N,E)$. The mapping in this case is done as follows:
 - (a) The specialized vocabulary V_s is the set of all distinct terms in the ICD-9 hierarchical alphabetical index excluding stop words. The generic vocabulary V_g is the set of all terms in the medical documents excluding stop words. The linguistic vocabulary is given by $V = V_s \cup V_g$.
 - (b) The set of classification codes C is composed of all the category and subcategory codes in the ICD-9 alphabetical index.
 - (c) The hierarchical coding structure CS-9 is built as follows: (i) each appearance of a term $t_i \in V_s$ in the ICD-9 alphabetical index is modeled as a node $n_i \in N$ unless it is a synonym term (indicated by a comma), (ii) a directed edge e_{ij} is created from a node n_i to a node n_j whenever $t(n_i)$ and $t(n_j)$ are consecutive terms in a codepath (see Section 2), (iii) a code c_k is inserted into the set C_{ij} whenever it is associated to a codepath which ends in $t(n_j)$ and has $t(n_i)$ as the last but one term, (iv) a synonym s_k is inserted into the set S_{ij} whenever it is associated to the term $t(n_j)$ (i.e., s_k appears immediately after $t(n_j)$ and they are separated by a comma) and is preceded by $t(n_i)$ in a codepath.
3. The acronym dictionary A and the synonym dictionary S are obtained from a specialist in assignment of ICD-9 codes.
4. The assignment of an code c to a medical document Q is modeled as a classification code assignment (Q, c, r) whose computation is detailed below.

Given that we have represented the ICD-9 codification problem in the framework of our hierarchical model, classification code assignments to a given medical document $Q = \{q_1, q_2, \dots, q_{nq}\}$ are computed as follows:

For each section $q_i \in Q$ do

1. substitute each acronym in the text of q_i by its respective term string;

2. traverse the text of q_i sequentially looking for terms associated to root nodes in CS-9;
3. for each root node found, determine the largest length path among all paths which satisfy the following condition: for each edge e_{ij} in the path, the terms $t(n_i)$ and $t(n_j)$ appear both in a same text window (belonging to the text of q_i) of size W centered around $t(n_i)$;
4. given the set of all largest paths found, let p_{min} be the smallest path length and p_{max} be the largest path length;
5. for each largest path P found of root node, let p be its path length and C_t be the set of codes associated to its terminal edge; then, to each code $c \in C_t$ generate the code classification assignment (Q, c, r) where the rank r is given by

$$r = \frac{p - p_{min}}{p_{max} - p_{min}} \times (1 - 0.8) + 0.8 \quad (2)$$

The adoption of a text window of size W ensures that paths traversed in our code hierarchy are formed by edges whose endpoint terms are closely related in the medical document (i.e., the endpoint terms are near one another in the text). For the test collection used in our experiments, a good value for W is 7. In the computation of the rank r , the threshold 0.8 is used to ensure that all ranks generated here are in the range $[0.8, 1.0]$. The reason is that, as discussed in Section 5, we also evaluate variations of our retrieval algorithm (which consider synonyms and approximate string matching) which are less precise and thus, retrieve a larger number of classification codes. To these codes, we assign ranks whose values are smaller than 0.8. Notice, however, that the threshold value 0.8 could be changed to 0.7 or some other value without affecting our results. All that is required is a separation in numerical ranges.

We observe that our computation of the r ranks is quite simple (the complexity is in finding the appropriate largest paths) and does not take into account characteristics of the terms in the coding path such as their *idf* factors. This is a next step which we plan to evaluate in the near future. However, despite the simplicity of the formula for the r ranks, our ranking computation procedure is quite effective as illustrated in our experiments.

5 Experimental Results

In this section, we describe a set of experiments which compare the retrieval performance of the ICD-9 code assignment algorithm based on the vector model and the ICD-9 code assignment algorithm based on the hierarchical model we proposed. For simplicity, we refer to the first one as the *vectorial* algorithm and to the second one as the *hierarchical* algorithm. The evaluation measures used are the standard precision and recall figures [2].

5.1 Test Collection and Characterization of the Experiments

Our test collection is composed of 2,424 ICD-9 distinct classification codes extracted from [6]. It does not include yet all classification codes in [6] due to the following reasons. First, we were unable to obtain a digital version of the ICD-9 document in Portuguese which forced us to digitize the text ourselves. Second, the digitization of the ICD-9 hierarchical alphabetical index required an OCR scanning phase

followed by a manual phase of syntax corrections which are both time consuming. Third, the Sarah Hospital Network deals mostly with orthopaedics and rehabilitation which implies that only a portion of the ICD-9 codes is required for classifying its medical documents. For the 2,424 codes we digitized, the largest codepath (see Section 2) has size 18 and the number of distinct terms in the specialized vocabulary is 4,700.

The reference queries used in our experiments are a set of 82 inpatient discharge summaries (in Portuguese) obtained from the Sarah Hospital in Belo Horizonte. Despite its small size, this set of reference queries was carefully selected from a much larger pool and, according to the specialists, is a representative sample of the discharge summaries produced by the hospital. To each of these summaries is assigned a set of category and subcategory codes (a task carried out by specialists in codification of the Sarah Hospital). We refer to the set of codes for a given inpatient discharge summary as its *ideal code set*.

Evaluation of our classification algorithms is done through recall and precision figures. These figures are generated by traversing the ordering of codes generated (by our algorithms) and verifying whether these codes are in the ideal code set. At any given point of this traversal, recall is the fraction of ideal codes which have been seen while precision is the fraction of traversed codes which are in the ideal code set.

Our set of experiments is characterized by three basic cases. In the first case, the hierarchical algorithm is compared with the vectorial algorithm. Two situations are considered: (a) take into account all queries and (b) take into account only the queries which return a non-empty code set. In the second case, approximate string matching is used for term matching both for the hierarchical and vectorial algorithms. In the third case, besides approximate string matching, synonyms are used with the hierarchical algorithm. In the immediately following, we discuss the effects of these variations in the retrieval performance of our algorithms and comment our results.

5.2 The Results

Figure 3 illustrates the curves of recall and precision for the vectorial and hierarchical algorithms described in Section 4. As in [3, 4], separate curves are plotted for the codification at the category and the subcategory levels. The results show that, for this set of queries, the hierarchical algorithm outperforms the vectorial algorithm for the codification at both the category and the subcategory levels. The reasons are twofold. First, the hierarchical algorithm takes advantage of the hierarchical topology of the index which is not done by the vectorial algorithm. Second, the largest path strategy adopted by the hierarchical ranking computation mimics a procedure which is common among specialists in code classification assignment.

We observed that 15% of the queries return an empty answer set and that another 10% of the queries return answer sets with no relevant codes (i.e., they yield 0% of precision at all recall levels). Furthermore, for the vectorial algorithm, such queries return poor answers. In Figure 4, we recompute the above precision and recall values considering only the queries which return a non-empty answer set for the hierarchical algorithm. The number of queries considered is now 69 i.e., there are 13 queries which return an empty answer set.

We observe that precision improves at all recall levels

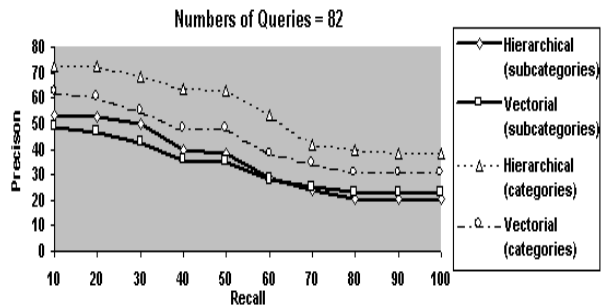


Figure 3: Basic hierarchical and vectorial algorithms. All queries considered (in number of 82).

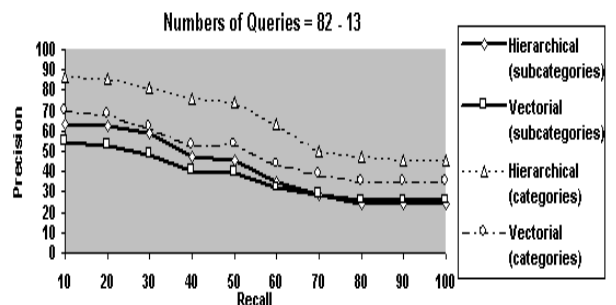


Figure 4: Precision and recall figures considering only queries with a non-empty answer set. Number of queries considered is now 69.

particularly for the category codes. Figure 5 illustrates the relative improvements obtained in precision figures. As observed, the improvement is as high as 22% at 30% of recall for subcategory coding and as high as 44% at 60% of recall for category coding.

We also notice that removing the queries with no answers from our set of test queries is reasonable because the algorithm might signal the specialist indicating which are those queries (and thus, which have to be classified entirely by the specialist). Thus, all of our remaining experiments consider only queries which return a non-empty answer set.

The results in Figure 5 can again be improved. We discussed this problem with the group of specialists in document codification at the Sarah Hospital and obtained from them the following insights. First, inpatient discharge summaries include misspellings and incomplete term specifications due to the short time doctors spend in writing them down. We deal with this situation through the adoption of a strategy of approximate string matching and through the introduction of synonyms as discussed below. Second, some summaries do not provide enough information for a proper codification. In these cases, the specialists discuss the case separately and might recur to documents other than the ICD-9 code structure. Further, the history of document codification in the medical institution might lead to the adop-

Recall	Subcategory			Category		
	Vectorial	Hierarchical	Improvement (%)	Vectorial	Hierarchical	Improvement (%)
10	54.73	63.21	15.49	69.95	86.35	23.45
20	52.57	62.37	18.64	67.66	85.85	26.88
30	48.51	59.23	22.10	61.48	81.62	32.76
40	40.22	47.18	17.30	53.09	75.72	42.63
50	39.72	45.65	14.93	53.24	74.50	39.93
60	31.74	34.51	8.73	43.84	63.24	44.25
70	28.38	28.35	-0.11	38.59	49.96	29.46
80	25.91	24.28	-6.29	35.19	46.98	33.50
90	25.80	24.28	-5.89	35.03	45.89	31.00
100	25.80	24.28	-5.89	35.05	45.89	30.93

Figure 5: Relative improvements in precision due to the hierarchical algorithm.

tion of coding rules which are in conflict with the ICD-9 coding strategy. To deal with this problem, we ask the specialists to review the ideal code set for each of our 82 queries according to the coding rules of the ICD-9 document. The results are discussed later on.

In our first variant, we modified our coding algorithms (both the hierarchical and the vectorial) to include an approximate string matching strategy. In this variant, a query term t_i retrieves all ICD-9 terms which differ from t_i by 0, 1, or 2 characters. As a result, more ICD-9 terms match any given query term. In our experiments, we used the software package Agrep [10] for performing approximate string matches. The computation of a rank r for a code c whose path length p includes terms with approximate matches (i.e., terms with mismatches of 1 or 2 characters) is now given by Equation 3 as follows

$$r = \frac{p - p_{min}}{p_{max} - p_{min}} \times (0.8 - 0.6) + 0.6 \quad (3)$$

This ensures that codes obtained using approximate term matching appear after the codes obtained using solely exact term matching. Figure 6 illustrates average precision and recall figures. Only queries with a non-empty answer set are considered.

From Figure 6, we first notice that the precision figures are lower than before both for category and subcategory coding. The reason is that the use of approximate term matching yields too many extra classification codes. The only advantage comes from the fact that more queries have now a non-empty answer set. In fact, the number of queries answered increased from 69 (Figure 4) to 72 (Figure 6). Despite this gain, the drop in average precision for the vectorial algorithm is excessive because this algorithm does not handle well the “noise” generated by the extra terms. Thus, in all our remaining experiments, we consider only the basic vectorial algorithm used in Figure 4. For the hierarchical algorithm, we continue to use approximate term matching because the drop in average precision is quite small and this drop in precision is compensated by the increment in the number of queries effectively answered.

In our second variant, we further modified the hierarchical algorithm to use synonyms (which, in this case, also include morphological variations such as feminine, masculine, and plural forms) in conjunction with approximate term matching. In this variant, a query term t_i also retrieves

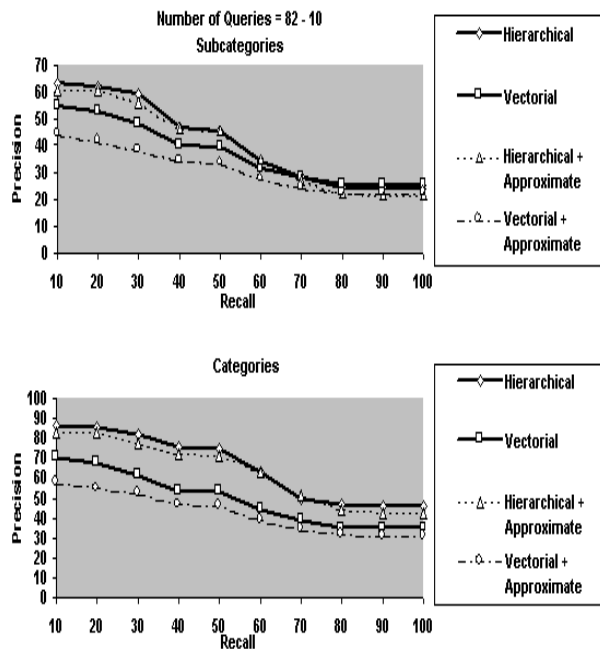


Figure 6: Precision and recall figures considering approximate string matching. Number of queries considered (only those with non-empty answer sets) is now 72.

all of its ICD-9 synonym terms. For the vectorial algorithm, the usage of synonyms does not yield good results and thus, as already commented, we stick to the vectorial algorithm of Figure 4. Figure 7 illustrates our results when synonyms are used in conjunction with approximate string matching.

From Figure 7, we observe an increment in average precision both for category and subcategory coding. We also observe that the number of queries with non-empty answer sets increases to 74. Furthermore, we included in Figure 7 the results for a heuristic variant of the hierarchical algorithm which we call *hier-approx-syn-near*. In this heuristic, besides adopting approximate term matching and synonyms, the algorithm also attempts to determine a nearby subcategory code whenever no subcategory code was found but a category code c_{cat} was. In this case, a nearby subcategory code is determined by looking at the set of all of subcategory codes associated to the category code c_{cat} and selecting one. The subcategory code selected is either the most frequently assigned one (a statistics obtained by observing the query set) or the default subcategory for the category code c_{cat} (whenever such default code exists in the ICD-9 alphabetical index). The results indicate an additional (small) gain in average precision for subcategory coding.

Finally, to counterbalance the fact that there are medical summaries which are incomplete (and yield no ICD-9 code) and the fact that the specialist sometimes adopts coding practices which violate ICD-9 coding rules, we asked our specialists to review the ideal code sets for our 82 queries. In this revision, they looked at all codes in each ideal code set and determined whether such code satisfied the ICD-9 coding rules. With the revised set of ideal code sets, we rerun our last set of experiments. The results are illustrated

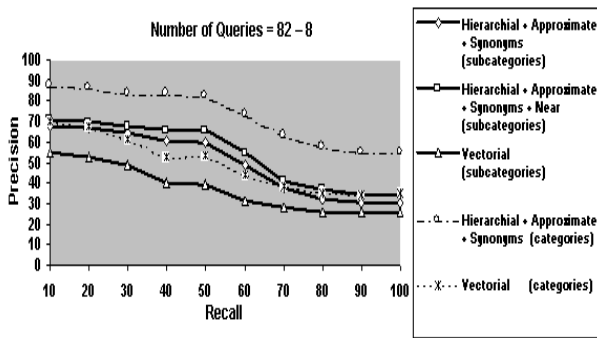


Figure 7: Precision and recall figures considering approximate string matching and synonyms. Number of queries considered is now 74.

in Figure 8.

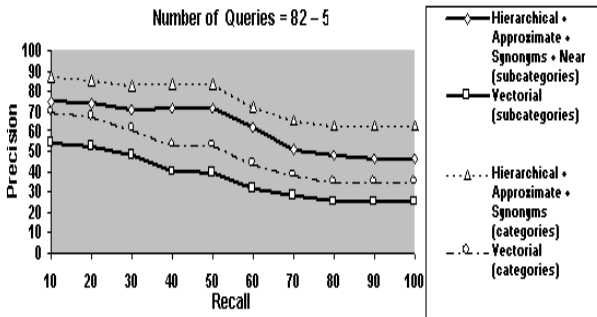


Figure 8: Precision and recall figures considering approximate string matching, synonyms, and a review of codes by the specialist. Number of queries considered is now 77.

As we can observe, average precision results improve after the review of the ideal code sets. Furthermore, the hierarchical algorithm presents a nice curve of average precision for subcategory coding and a very nice curve of average precision for category coding. Figure 9 illustrates the relative improvements in precision obtained by the hierarchical algorithm. We notice that average precision at high recall is now quite good and that the relative improvements in precision at high recall are considerable (both for category and subcategory coding). Furthermore, the number of queries answered is now 77 (out of 82).

6 Conclusions

In this paper, we described a hierarchical model we proposed for the automatic categorization of medical documents, and presented a comparative analysis of two algorithms for automatic assignment of ICD-9 classification codes. The first algorithm is based on the classic vector space model, while the second one is based on our hierarchical model and takes advantage of the hierarchical structure of the ICD coding scheme. The analysis is based on experiments carried out

Recall	Subcategory			Category		
	Vectorial	Hierarchical	Improvement (%)	Vectorial	Hierarchical	Improvement (%)
10	61.14	80.74	32.06	74.47	91.24	22.52
20	56.90	79.09	39.00	70.18	89.74	27.87
30	49.72	74.80	50.44	62.92	87.09	38.41
40	42.02	74.45	77.18	53.64	87.71	63.52
50	41.89	74.35	77.49	53.82	87.89	63.30
60	32.31	64.05	98.24	43.91	75.93	72.92
70	28.89	51.11	76.91	39.09	68.53	75.31
80	26.49	50.96	92.00	35.50	66.60	87.61
90	25.27	49.91	97.51	35.05	66.47	89.64
100	25.27	49.91	97.51	35.06	66.47	89.59

Figure 9: Relative improvements in precision obtained by hierarchical algorithm (includes approximate string matching and synonyms) with revised ideal code sets. Number of queries considered is now 77.

using a medical database from a Brazilian hospital. Our results show that the algorithm based on our hierarchical model outperforms the algorithm based on the classic vector space model and that it yields quite good precision and recall figures for the problem of ICD-9 code assignment. The reasons are twofold: (1) our model incorporates the hierarchical structure of the ICD-9 index and (2) it captures knowledge from the coding specialist. It is also important to point out that, despite being based on the ICD, our model is quite general and is equally applicable to other hierarchical medical coding standards.

In the near future, we plan to extend our study in two directions. First, we intend to carry out new experiments introducing other components proposed in our model (e.g., a vocabulary of generic medical terms [5]). Second, we are implementing an interactive Web interface for ICD-9 code assignment that will be used at the Sarah Hospital to assess the impact of our approach on a real environment.

Acknowledgments

We would like to thank Andréa Faria and Tânia Rozalina, coding specialists of the Sarah Hospital in Belo Horizonte, for their advice on the ICD-9 code assignment problem. Thanks are also due to Gustavo Mendonça for helping us with the implementation of the vectorial algorithm. Data used in our experiments have been made available to us by the Sarah Hospital Network.

References

- [1] J.J. Cimino. Vocabulary and health care information technology: State of the art. *Journal of the American Society for Information Science*, 46(10):777–782, 1995.
- [2] W.B. Frakes and editors R. Baeza-Yates. *Information Retrieval: Data Structures & Algorithms*. Prentice Hall, Englewood Cliffs, New Jersey, 1992.
- [3] L.S. Larkey and W.B. Croft. Automatic assignment of ICD9 codes to discharge summaries. Technical report, Center for Intelligent Information Retrieval at University of Massachusetts, Amherst, Massachusetts, 1995.

- [4] L.S. Larkey and W.B. Croft. Combining classifiers in text categorization. In *Proceedings ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 289–297, Zurich, Switzerland, 1996.
- [5] L.R.S. Lima, A.H.F. Laender, and A.B. Ribeiro-Neto. A specialized model for information retrieval applied to semi-structured medical databases. In *Proceedings of XII Brazilian Symposium on Database Systems - SBBD'97*, pages 241–256, Fortaleza, Brazil, 1997. (in Portuguese).
- [6] Organização Pan-Americana de Saúde. *Classificação Internacional de Doenças, Revisão 9*. EDUSP - Editora Universidade de São Paulo, São Paulo, Brazil, 1980. (Volumes 1 e 2).
- [7] J. Pearl. *Probabilistic Reasoning in Intelligent System: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, California, 1988.
- [8] G. Salton and C. Buckley. Term-weighting approaches in automatic retrieval. *Information Processing & Management*, 24(5):513–523, 1988.
- [9] Y. Satomura and M.B. Amaral. Automated diagnostic indexing by natural language processing. *Medical Informatics*, 17(3):149–163, 1992.
- [10] S. Wu and U. Manber. Fast text searching allowing errors. *Communications of the ACM*, 35(10):83–91, 1991.