

Improving Text Retrieval in Medical Collections Through Automatic Categorization

Authors omitted for blind review

No Institute Given

Abstract. A current and important research issue is the retrieval of relevant medical information. In fact, while the medical knowledge expands at a rate never observed before, its diffusion is slow. One of the main reasons is the difficulty in locating the relevant information in the modern and large medical text collections of today. In this work, we introduce a framework, based on Bayesian networks, that allows combining information derived from the text of the medical documents with information on the diseases related to these documents (obtained from an automatic categorization method). This leads to a new ranking formula which we evaluate using a medical reference collection (the OHSUMED collection). Our results indicate that this combination of evidences might yield considerable gains in retrieval performance. When the queries are strongly related to diseases, these gains might be as high as 84%. This shows that information generated by an automatic categorization procedure can be used effectively to improve the quality of the answers provided by an information retrieval (IR) system specialized in the medical domain.

1 Introduction

Today, we observe the development of the concept of medical informatics, which presents itself as a new area in the medical field. In the broad sense, this concept is also associated with the improvement of the tasks of searching, synthesizing, organizing, and disseminating medical information to doctors, patients, and people interested in health related issues in general. These considerations indicate the great importance of combining knowledge from the medical and information sciences fields into modern information retrieval (IR) systems.

The biomedical literature grows at a rate of 6% to 7% a year, doubling in size every 10 to 15 years. Most part of this literature is now available in some electronic form and frequently accessible through the Internet. Such availability facilitates the access to specialized medical information, but introduces new problems in its own. While the medical knowledge expands at a rate never observed before, its diffusion is slow. The barriers for the diffusion of the medical knowledge are many and include: the limited time for bibliographical searching, the limited access to the information sources, and the great difficulty of doctors and others medical professionals in identifying the relevant information within the vast medical collections of today [13]. In this work, we focus on this last issue

i.e., on the problem of improving the quality of the answers returned to queries focussed on medical topics.

A standard approach to this problem is to apply standard information retrieval (IR) techniques to the medical domain. While this approach does provide a solution to the problem of finding relevant information in a large medical collection, it does not take into account any specialized information from the medical arena. This clearly seems to be a strong limitation.

An alternative approach is to develop a framework that allows combining IR techniques with knowledge from the medical domain. This is the path we follow here. We consider a specific form of medical knowledge, i.e., information on the diseases related to the documents in a medical collection. Given a medical collection, information on diseases can be generated through the assignment of ICD (International Code of Diseases) codes to the documents of the collection. This can be accomplished in fully automatic mode with great effectiveness, as we later discuss. Given the information on the diseases related to the medical documents (through ICD codes), we study the problem of how to improve the quality of the answers generated (i.e., how to improve the retrieval performance of the system).

To combine information on ICD codes with information derived from the text of the documents (which is the information used by the standard IR ranking algorithms), we adopt the framework of Bayesian networks [15]. Bayesian networks are useful because they allow combining distinct sources of evidence in consistent fashion. Also, they provide an intuitive modeling tool that facilitates capturing (in the model) the influence of the key parameters of the problem being modeled. The Bayesian framework we adopt leads to a new ranking formula that takes into account information about the text of the medical documents and information about the diseases related to the documents. Through experimentation, we show that this leads to improved retrieval performance. When only queries that are strongly related to diseases are considered, the average improvement in retrieval performance is as high as 84% (compared to a standard IR algorithm).

The paper is organized as follows. Section 2 discusses related work. Section 3 briefly describes a method for automatically categorizing medical documents we developed and that we use to assign ICD codes to the documents of the collection. Section 4 describes our ranking function based on the Belief network model. In Section 5 we discuss our experimental results which are based on the OHSUMED reference collection. In Section 6 we present our conclusions.

2 Related Work

Several approaches have been proposed in the literature to address the problem of automatic categorization of medical documents. In [9,10], the problem is treated as a classification problem which is solved by combining three classifiers based on probabilistic models [12] for the code assignment task. In [18], the problem is addressed through the usage of natural language processing techniques. Both approaches present good results in some situations but have some

disadvantages. 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, in both approaches the hierarchical structure of the coding standard and the knowledge of the coding specialists are completely ignored. In [?,?] ¹, a method for automatic categorization of medical documents that takes advantage of the hierarchical topology of coding schemes such as the International Code of Diseases (ICD) to generate high precision results.

Automatic text categorization has been used in different applications such as text classification, text filtering, and text retrieval. Applications in text retrieval, in particular, have received special attention. Methods such as decision trees [1], linear classifiers [11], context-sensitive learning [5], and learning by combining classifiers [10] have been proposed to address this problem.

Yang and Chute propose a method, known as Linear Least Square Fit (LLSF) [22], to perform automatic text categorization and text retrieval. The LLSF method uses a training set of manually categorized documents to learn word-category associations which are then applied to predict the categories of arbitrary documents. Similarly, this method uses a training set of queries and their related documents to obtain empirical associations between query words and indexing terms of documents, and then applies these associations to predict the related documents of arbitrary queries. Another method, called Expert Network [21], also needs a training set to categorize information. According to this method, the terms in a document are linked with its categories by means of a network in which there is a weight in each link. However, this method is preferable because it is simpler and computationally more efficient than the LLSF method.

Another approach, proposed in [8], also consists in a method for automatic categorization and a method for text retrieval. The categorization method derives from a machine-learning paradigm known as instance-based learning and an advanced document retrieval technique known as retrieval feedback. The text retrieval method computes two rankings: one for the free-text portion of the documents and another one for the category portion of the documents. The categories are generated according to the categorization method. The method proposes to sum both rankings and control the relative emphasis of each one through a parameter.

Our work is related to these approaches but we use different techniques for automatic categorization and text retrieval. The automatic categorization is supported by the method we proposed [?,?] and we use Bayesian networks [15, 20] to merge the rankings generated for the free-text and category portions of the documents. Bayesian networks supply the formalism for representing, quantifying, and combining two or more sources of evidence to support a ranking for the documents in the answer set. In this work, we use this method to represent and

¹ References removed for blind review

combine concept-based and text-based evidential information in a similar way as discussed in [15, 19].

In spite of the differences in the above related works, all of them demonstrate that the categorization process, being it automatic or manual, is effective because it can be used to improve the retrieval performance when compared with a system that uses no categorization.

3 The Automatic Categorization Method

In this section, we briefly describes our method for automatic assignment of ICD-9 codes to medical documents. The method uses the hierarchical structure of the ICD-9 alphabetical index to guide the coding task and attains levels of precision comparable to the codification provided by medical specialists [14]. While in here we focus on the ICD-9 alphabetical index, we notice that the method can be used with other medical coding schemes such as SNOMED and MeSH.

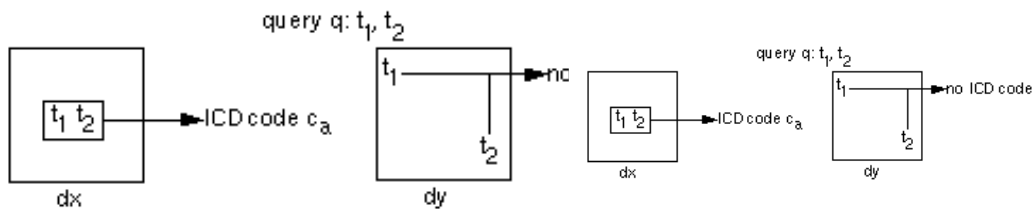


Fig. 1. An example of the terms hierarchy for the ICD-9 vocabulary.

Fig. 2. Example of the categorization process.

Fig. 1 shows some fragments of the ICD-9 alphabetic index. As we can see, the code 001.9 is associated with the term *Cholera*, the code 985.9 is associated with the sequence of terms *Cholera + Antimonial*, and so on.

Fig. 2 illustrates the categorization process. This process can be represented by a function $codes(d)$ that, given an input document d , returns a list of codes C related to d . This function scans the document d searching for terms or sequence of terms that lead to codes in the ICD-9 alphabetic index. For instance, for the document in Fig. 2 the function $codes(d)$ returns the codes 276.2, 001.9, 273.8 and 276.2 which are related, respectively, to the terms Acidosis, Cholera, hyperproteinemia, and acidemia. For more details on our categorization method we refer the reader to [?,?].

4 The Ranking Fusion Model

In this work, one of our main goals is to investigate whether knowledge derived from the diseases associated with a medical document (i.e., information about

its ICD codes) can be used to improve retrieval performance. Our approach is to combine evidence from the vector model with evidence from the ICD categorization and to investigate whether gains in retrieval performance can be attained.

To combine these two sources of evidence, we use a framework based on Bayesian networks. Bayesian networks are useful because they allow combining distinct sources of evidence in a consistent fashion and also provide an intuitive modeling tool that facilitates capturing (in the model) the influence of the key parameters of the system. Further, they have been used successfully with various reference collections and for distinct purposes in the past [3, 4, 6, 15, 19, 20].

In here, we extend the evidence provided by the classic vector model in the belief network [16, 19], to include evidence from ICD categories. Fig. 3 illustrates the extended network.

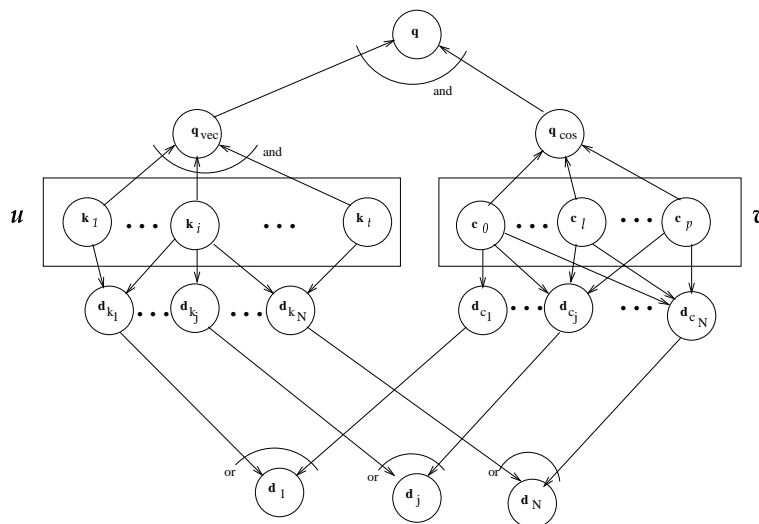


Fig. 3. Bayesian network expanded with evidence from ICD categories.

The right part of the network models ICD codes (represented by the c_i nodes) and their relationships with the query (represented by the q_{cos} node) and with the medical documents (represented by the d_{c_i} nodes). In the left part of the network, the query node is renamed as q_{vec} and the documents nodes are renamed as d_{k_i} to allow distinguishing between the representations for the query and for the documents in the right and left parts of the network. An extra node q is inserted at the top of the network to represent the fact that the query now considers evidence from the vector model (through the q_{vec} node) and evidence from the ICD categories (through the q_{cos} node). Extra nodes d_j are inserted at the bottom of the network to represent the fact that a document d_j now considers evidence from the vector model (through the dk_j node) or evidence from the ICD categories (through the dc_j node). Notice that evidence at the

query side of the network is combined through a conjunctive operator, while evidence at the document side of the network is combined through a disjunctive operator. This works better as discussed in [15, 19].

Let u represent the state of the set of the k_i root nodes, and let v represent the state of the set of c_i root nodes. For the k_i root nodes, we consider only the states u_i such that

$$u = u_i \iff g_i(u) = 1 \wedge g_{j \neq i}(u) = 0$$

In Fig. 3, the rank $P(d_j|q)$ associated with a document d_j is computed through basic conditioning on the root nodes and application of Bayes' rule as follows.

$$\begin{aligned} P(d_j|q) &= \frac{P(d_j \wedge q)}{P(q)} \\ &= \eta \sum_{u,v} P(d_j|u,v) P(q|u,v) P(u) P(v) \\ &= \eta \sum_{u,v} [1 - (\overline{P(dk_j|u)}) (\overline{P(dc_j|v)})] P(q_{vec}|u) P(q_{cos}|v) P(u) P(v) \quad (1) \end{aligned}$$

where η is a normalizing constant.

Consider the situation in which two documents, d_x and d_y , contain exactly the same set of query terms. Assume that the query terms in document d_x lead to an ICD code and that this does not occur for document d_y (i.e., no codes are assigned to document d_y). Fig. 4 illustrates a situation in which this might happen. In this case, we expect the document d_x to have a combined final ranking that is higher than the ranking for d_y (because $P(d_x|v) > P(d_y|v)$).

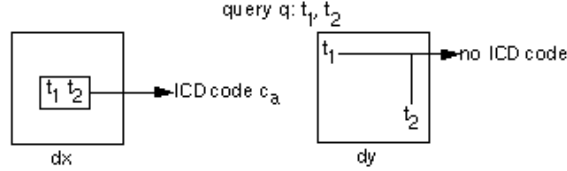


Fig. 4. Two documents, d_x and d_y , contain the query terms but only one of them d_x leads to an ICD code.

However, this is not guaranteed by Equation (1) because $P(d_x|u)$ and $P(d_y|u)$ might differ due to: (a) the frequencies of t_1 and t_2 in the two documents might differ and (b) the normalization factors $|d_x|$ and $|d_y|$ might also differ. To avoid these side effects, we adopt a ranking that consider only the *idf* factor of the classic vector model [17] at the left side of the network. As a result, we define the probabilities $P(q_{vec}|u)$ and $P(dk_j|u)$ as follows:

$$P(q_{vec}|u) = \begin{cases} 1 & \text{if } u = u_{qvec} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$P(dk_j|u) = \begin{cases} \sum_{k_i \in u_{qvec}} \frac{\log \frac{N}{n_i}}{|d\mathbf{k}_{max}|} \times \frac{\log \frac{N}{n_i}}{|q_{vec\ max}|} & \text{if } u = u_{qvec} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where n_i is the number of documents in which the keyword k_i appears. N is the total number of documents in the collection.

Let us now turn our attention to the right side of the network. The probability $P(q_{cos}|v)$ quantifies the relationship between the ICD categories and the query q . Larger the coverage of the query concept q provided by the ICD categories, more related to diseases the query is. This is important because we should not expect gains in retrieval performance (i.e., in the quality of the ranking) due to ICD categories, if the query is not related to diseases. To quantify this coverage relationship, we use the number of terms in common between the query q and the ICD codes, as follows.

Let $codes(q)$ be a function that returns the set of codes generated by our coding method for the query q . Consider the 2^p possible states for the set v of root nodes. Instead of using the states in which a single node c_l is active at a time (as done for the left side of the network), we use only a single state that includes all c_l codes in the set $codes(q)$. We do so because this simplifies the computation of the coverage relationship. Define the state v_q of the set v of root nodes, as follows:

$$v = v_q \text{ iff } \begin{cases} g_l(v) = 1 \ \forall l | c_l \in codes(q) \\ g_l(v) = 0 \text{ otherwise} \end{cases} \quad (4)$$

Equation (4) defines v_q as the state of v that contains the nodes $c_l \in codes(q)$ active and the nodes $c_l \notin codes(q)$ inactive.

For each $c_l \in codes(q)$, let \mathbf{c}_l be a vector of binary term weights, where each term weight is 1 to indicate that term is associated with the code c_l (according to the ICD hierarchy), and 0 otherwise. Also, let \mathbf{q}_{cos} be a vector of binary term weights, where each term weight is 1 to indicate that the term occurs in the query q , and 0 otherwise. The product $\mathbf{c}_l \bullet \mathbf{q}_{cos}$ provides a measure of the coverage relationship between the concepts c_l and q_{cos} . We intend to identify the code c_l that best covers the query q (and thus, which is more likely to define the central disease associated with the query q). To accomplish this effect, we define:

$$P(q_{cos}|v) = \begin{cases} \max_{\forall l | g_l(v)=1} \frac{\mathbf{c}_l \bullet \mathbf{q}_{cos}}{|\mathbf{c}_{max}| \times |\mathbf{q}_{cos}|} & \text{if } v = v_q \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$$P(\bar{q}_{cos}|v) = 1 - P(q_{cos}|v)$$

Notice, that we use $|\mathbf{c}_{max}|$, instead of $|c_l|$, because we want to measure the coverage relationship with regard to the query q only.

For the probabilities $P(d_j|v)$, we are interested in a slightly distinct form of coverage relationship. Instead of simply looking at common terms, we look at the coverage relationship between the codes associated with the query q and the codes associated with a document d_j . This is an important point because a code c_l , $c_l \in codes(q)$, might have terms in common with a document d_j even if this code is not associated with d_j (as illustrated in Fig. 4). Thus, we must focus on the coverage relationship between $codes(q)$ and $codes(d_j)$. For this, we define:

\mathbf{v}_q : vector of code weights associated with $c_l \in codes(q)$,
 \mathbf{d}_{c_j} : vector of code weights associated with $c_l \in codes(d_j)$;

The code weights here do not take into account code frequencies but do have an *idf* component (computed over the set of all codes assigned to all documents in the collection). This leads to a ranking form which yields:

$$P(dc_j|v) = \begin{cases} \sum_{c_i \in q_{cos}} \frac{\log \frac{N}{C_i}}{|\mathbf{d}_{c_{max}}|} \times \frac{\log \frac{N}{C_i}}{|\mathbf{q}_{cos_{max}}|} & \text{if } v = v_{q_{cos}} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

$$P(\overline{dc_j}|v) = 1 - P(dc_j|v)$$

where C_i is the number of documents in which the code c_i appears. N is the total number of documents in the collection.

As a result of the definition of $P(q_{vec}|u)$, we have:

$$P(d_j|q) = \eta [1 - (\overline{P(dk_j|u_{qvec})}) (\overline{P(dc_j|v_{qcos})})] P(q_{cos}|v) P(u) P(v) \quad (7)$$

Finally, the prior probabilities $P(v)$ and $P(u)$ are set to constants.

5 Experimental Results

We first present the medical reference collection we used in our experiments. Following, we discuss our results.

5.1 The Medical Reference Collection

The reference collection used in our experiments was the OHSUMED collection [7] which has been widely used for experimentation in the medical domain. The OHSUMED collection contains 348,566 references, which are derived from the subset of 270 journals found in the KF MEDLINE Primary Care session, covering the years from 1987 to 1991. The collection includes 106 example queries that were generated by actual physicians in the course of patient care. For each example query, at least one definitely relevant document is indicated. Each query is formed by a brief statement about the patient, followed by a description of the information need. The collection also includes relevance judgments for the example queries. Each relevance judgment indicates a document as definitely

relevant, possibly relevant, or irrelevant. In our experiments, we used only documents with an abstract. This generated a new subcollection having 233,445 documents and 93 queries with relevant documents.

Fig. 5 quantifies the relationship between the ICD categories and each of our 93 test queries, according to Equation (13). This relationship indicates the support provided to each query by the ICD categories and is here referred to as the “query icd relation factor” (or simply, icd-relation-factor). As we can see, 14 of such queries are not related to diseases (icd-relation-factor = 0) and 69 queries have some relationship with diseases (icd-relation-factor > 0). From these 69 disease-related queries, 55 have a good focus on diseases (icd-relation-factor ≥ 0.5) and 10 are highly related to diseases (icd-relation-factor ≥ 0.8). The 69 disease-related queries are the focus of our experiments.

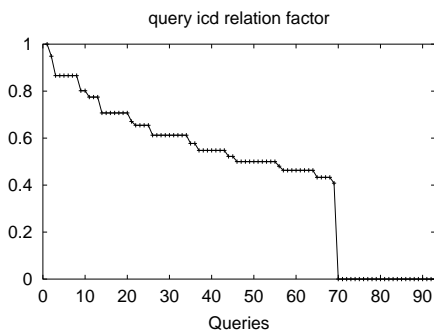


Fig. 5. Variation of the icd-relation-factor for our 93 test queries. The queries are sorted by decreasing values of the icd-relation-factor.

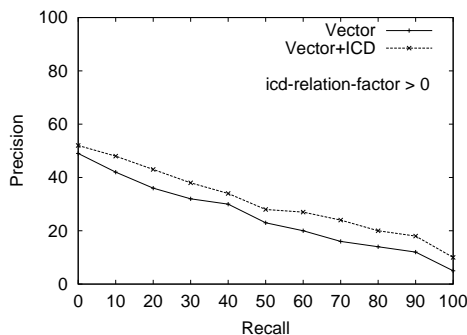


Fig. 6. Average precision figures for the vector and vector+ICD rankings. Only the 69 disease-related queries are considered.

5.2 Results

In this section, we present our experimental results. Fig. 6 illustrates the retrieval performance, in terms of precision-recall figures [2], for the vector and vector+ICD rankings in our belief network model. We consider only the 69 example queries that are effectively related to diseases (i.e., those queries for which icd-relation-factor > 0). This is appropriate because we should not expect gains in retrieval performance due to evidence from ICD categories, whenever the query is not related to diseases.

In Fig. 6, we observe that the vector+ICD ranking is always superior for our reference collection. Table 1 details these results, which show that adding a new source of evidence, based on the ICD categorization, to the vector-based evidence leads to superior results. Another important observation is that the average recall (computed over all queries) for the vector+ICD model is 7% higher than for the vector model. This is because the ICD categories allow finding new documents,

which are related to the user query, and which are not retrieved when only the term evidence is used.

<i>Average precision figures for 69 queries(icd-relation-factor > 0)</i>			
Recall	Vector	Vector+ICD	Gain
0%	49.28	52.39	06.31%
10%	42.68	48.82	14.38%
20%	36.05	43.74	21.34%
30%	32.66	38.95	19.27%
40%	30.82	34.09	10.63%
50%	23.54	28.57	21.39%
60%	20.67	27.64	33.75%
70%	16.75	24.69	47.45%
80%	14.89	20.74	39.34%
90%	12.72	18.37	44.42%
100%	05.55	10.85	95.63%
Average	25.96	31.72	22.16%

Table 1. Average precision figures for the 69 disease-related queries.

Let us now focus on the queries that are more closely related to diseases. Fig. 7 illustrates the retrieval performance of the vector and vector+ICD rankings, when only queries with an icd-relation-factor \geq than 0.5 are considered (i.e., 55 example queries). Table 2 details these results. Again we observe that the vector+ICD ranking always yields higher precision figures than those obtained by the vector model. Further, the relative gain in precision is higher for these 55 queries than for the complete set of 69 disease-related queries. The reason for this better result is the higher relevance of the diseases for these 55 queries. This suggests that more related to diseases a query is, higher is the improvement provided by our extended network ranking. Our following set of results confirms this interpretation.

Fig. 8 provides precision-recall figures for the 10 queries with an icd-relation-factor \geq 0.8. Table 3 details these results. For these 10 queries, the vector+icd ranking yields a gain of 84,92% in the average precision, relative to the vector ranking. The average recall for the vector+ICD model is now 12,5% higher than for the vector model. Table 4 shows the average recall figures for each of our 3 query pools (selected by the icd-relation-factor).

6 Conclusions

We have described a framework for combining evidence derived from the text of medical documents with evidence provided by diseases related to these documents. The information on diseases is generated by a fully automatic catego-

Recall	Vector	Vector+ICD	Gain
0%	38.78	49.98	28.89%
10%	35.33	47.71	35.06%
20%	31.20	43.09	38.13%
30%	29.27	37.14	26.89%
40%	25.46	30.68	20.51%
50%	21.52	26.66	23.93%
60%	20.34	25.76	26.66%
70%	17.87	23.61	32.12%
80%	15.47	19.91	28.72%
90%	14.37	18.32	27.47%
100%	07.82	10.34	32.28%
Average	07.82	30.29	29.45%

Table 2. Average precision figures for the 55 queries (icd-relation-factor ≥ 0.5)

Recall	Vector	Vector+ICD	Gain
0%	30.52	49.91	63.56%
10%	22.77	53.75	136.02%
20%	25.35	51.29	102.29%
30%	18.39	42.10	128.88%
40%	20.45	37.46	83.20%
50%	17.60	32.08	82.25%
60%	17.25	29.64	71.80%
70%	16.16	28.49	76.37%
80%	15.08	24.95	65.46%
90%	15.13	20.71	36.89%
100%	09.84	15.22	54.68%
Average	18.96	35.05	84.90%

Table 3. Average precision figures for the 10 queries with icd-relation-factor ≥ 0.8 .

icd-relation-factor	Vector	Vector+ICD	Gain
> 0.0	82.60%	88.40%	7.00%
≥ 0.5	81.81%	87.27%	6.67%
≥ 0.8	80.00%	90.00%	12.50%

Table 4. Average recall for each query pool.

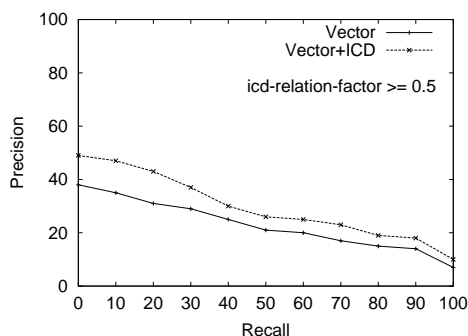


Fig. 7. Average precision figures for the vector and vector+ICD rankings, restricted to the 55 queries for which $\text{icd-relation-factor} \geq 0.5$.

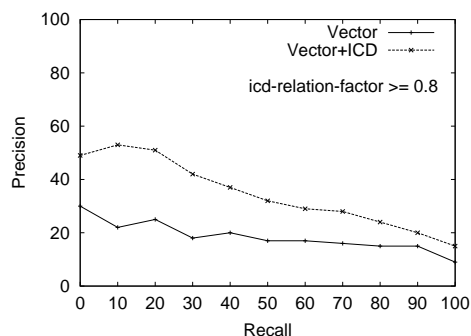


Fig. 8. Average precision figure for the vector and vector+ICD rankings, restricted to the 10 queries for which $\text{icd-relation-factor} \geq 0.8$.

rization method we developed, which assigns ICD codes to the documents in a medical collection.

Our framework is based on Bayesian networks. Bayesian networks are useful because they allow combining distinct sources of evidence in a consistent fashion. The Bayesian framework we proposed leads to a new ranking formula that takes into account information about the text of the medical documents and information about the diseases related to these documents. Through experimentation with a medical reference collection (the OHSUMED collection), we evaluated the effectiveness of our approach. We considered 3 distinct pools of test queries: queries that mention diseases, queries related to diseases, and queries focused on diseases. In all three cases, our new ranking formula yielded improved retrieval performance when compared to a standard IR ranking algorithm (the vector model, which we have adopted as our baseline). When only queries that are strongly related to diseases were considered, the average improvement in retrieval performance was as high as 84%. Our results show the importance of taking into account specialized medical information in medical retrieval systems.

Besides providing improved retrieval performance, our method for the automatic assignment of ICD codes generates a categorization hierarchy that includes more than 5,000 diseases (those included in the ICD hierarchy). This is a fine hierarchy which aggregates knowledge to large medical collections such as the Medline. This hierarchy can be used, for instance, to naturally design an interface based on a large directory of diseases. In the near future, we intend to experiment with such a hierarchy and evaluate its effectiveness in facilitating the access of medical information of relevance.

References

1. C. Apte, F. Damerau, and S. M. Weiss. Automated Learning of Decision Rules for Text Categorization. *ACM Transactions on Information Systems*, 12(3):233–251, 1994.
2. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley Longman, Harlow, England, 1999.
3. J. Broglio, J.P. Callan, W.B. Croft, and D.W. Nachbar. Document retrieval and routing using the inquiry system. In *Proceedings of the Third Text Retrieval Conference - TREC-3*, pages 241–256, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 1995. (NIST Special Publication 500-225).
4. J. Callan. Document filtering with inference networks. In *Proc. of the 19th ACM-SIGIR Conference*, pages 262–269, Zurich, Switzerland, 1996.
5. W.W. Cohen and Y. Singer. Context-sensitive Learning Methods for Text Categorization. In *Proc. 19th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pages 307–315, Zurich, Switzerland, 1996.
6. D. Haines and W.B. Croft. Relevance feedback and inference networks. In *Proc of the Sixteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 2–11, Pittsburgh, PA, USA, 1993.
7. W. Hersh, C. Buckley, T. Leone, and D. Hickam. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proc. of 17th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 192–201, Dublin, Ireland, 1994. ACM.
8. W. Lam, M. Ruiz, and P. Srinivasan. Automatic Text Categorization and its Application to Text Retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 11(6):865–879, 1999.
9. 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.
10. L. S. Larkey and W. B. Croft. Combining Classifiers in Text Categorization. In *Proc. 19th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pages 289–297, Zurich, Switzerland, 1996.
11. D. D. Lewis, R. E Schapire, J.P. Callan, and R. Papka. Training Algorithms for Linear Text Classifiers. In *Proc. 19th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pages 298–306, Zurich, Switzerland, 1996.
12. J. Pearl. *Probabilistic Reasoning in Intelligent System: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, California, 1988.
13. S. L. Pestotnik. Medical informatics: Meeting the information challenges of a changing health care system. *Journal of Informed Pharmacotherapy*, 2(1), 2000.
14. B. Ribeiro-Neto, A.H.F. Laender, and L.R.S. Lima. An experimental study in automatically categorizing medical documents. *Journal of the American Society for Information Science and Technology*, 52(5):391–401, 2001.
15. B. Ribeiro-Neto and R. Muntz. A Belief Network Model for ir. In *Proceedings of the XIX ACM SIGIR International Conference on Information Retrieval*, pages 253–260, Zurich, Switzerland, 1996.
16. B. Ribeiro-Neto, I. Silva, and R. Muntz. Bayesian network models for information retrieval. In *In: Soft Computing in Information Retrieval*, pages 259–291, Physica-Verlag, Heidelberg, 2000. F. Crestani & G. Pasi, editors.
17. G. Salton and C. Buckley. Term-weighting approaches in automatic retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

18. Y. Satomura and M.B. Amaral. Automated diagnostic indexing by natural language processing. *Medical Informatics*, 17(3):149–163, 1992.
19. I. Silva, B. Ribeiro-Neto, P. Calado, E. Moura, and N. Ziviani. Link-based and Content-based Evidential Information in a Belief Network Model. In *ACM SIGIR 23rd Int. Conference on Information Retrieval*, pages 96–103, Athens, Greece, 2000.
20. H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, July 1991.
21. Y. Yang. Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval. In *Proc. 17th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pages 13–22, 1994.
22. Y. Yang and C. Chute. An Application of Least Squares Fit Mapping to Text Information Retrieval. In *Proc. 16th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pages 281–290, 1993.