# Multi-Classification of Clinical Guidelines in Concept Hierarchies

**Diego Sona**[1], **Paolo Avesani**[1] **and Robert Moskovitch**[2]
[1] ITC/irst - Trento, Italy
{sona,avesani}@itc.it
[2] Ben Gurion University - Beer Sheva, Israel
robertmo@bgumail.bgu.ac.il

## Abstract

Clinical practice guidelines (CPGs) are increasingly common in clinical medicine for prescribing a set of rules that a physician should follow. Recent interest is in accurate retrieval of CPGs at the point of care. Examples are the CPGs digital libraries National Guideline Clearinghouse (NGC) or Vaidurya (DeGeL), which are organized along predefined concept hierarchies, like MeSH and UMLS. In this case, both browsing and concept-based search can be applied. Mandatory step in enabling both ways to CPGs retrieval is manual classification of CPGs along the concepts hierarchy. This task is extremely time consuming. Supervised learning approaches, where a classifier is trained based on a meaningful set of labeled examples is not a satisfying solution, because usually too few or no CPGs are provided as training set for each class. In this paper we present how to apply the *TaxSOM* model for multi-classification. *TaxSOM* is an unsupervised technique that supports the physician in the classification of CPGs along the concepts hierarchy, even when no labeled examples are available. This model exploits lexical and topological information on the hierarchy to elaborate a classification hypothesis for any given CPG. We argue that such a kind of unsupervised classification can support a physician to classify CPGs by recommending the most probable classes. An experimental evaluation on various concept hierarchies with hundreds of CPGs and categories provides the empirical evidence of the proposed technique.

## 1 Introduction

Clinical practice guidelines (CPGs) are an increasingly common and important format in clinical medicine for prescribing a set of rules and policies that a physician should follow. According to studies, clinical guidelines improve medical practice. They improve the quality (and possibly also the cost-efficiency) of care in an increasingly complex health care environment [Grimshaw and Russel, 1993]. It would be best if automated support could be offered to guideline-based care at the point of care.

To support tasks such as the run-time application of a guideline, it is often important to be able to quickly retrieve a set of guidelines most appropriate for a particular patient or task. Correctly classifying the guidelines, along as many semantic categories as relevant (e.g., therapy modes, disorder types, sighs and symptoms), supports easier and more accurate retrieval of the relevant guidelines using concept based search. This approach is implemented in *Vaidurya* – a concept based and context sensitive search engine for clinical guidelines [Moskovitch *et al.*, 2004], which is the search engine of the *Digital Electronic GuidelinE Library* (DeGeL). Electronic CPG repository, such as the National Guideline Clearinghouse (NGC) provide a hierarchical access to electronic CPGs in a free-text or semi-structured format (see <http://www.ngc.org>).

The construction of such concept hierarchies and the consequent classification of CPGs along the provided concepts is usually committed to physicians that in the following of this paper will be also referred to as *"taxonomy editors"*. This classification, however, is mostly manual and extremely time consuming. Thus, an automatic process where CPGs are classified automatically along the concepts hierarchy is crucial, while very challenging.

The main aim of this paper is to provide a tool that assists the domain expert (physician), who classifies the CPGs. The idea is that whenever the physician needs to classify a set of CPGs, the tool provides recommendations on the most probable classes for each CPG. In particular, the tool is specially suited to help the physician when concept hierarchy is built from scratch, and no examples of labeled CPGs are provided for each class. In this case there is not any premise for a successful training of any existing supervised classifier, therefore, recommendations can be given only using an unsupervised model. We refer this task to as the *bootstrapping* problem [McCallum and Nigam, 1999; Adami *et al.*, 2003a]. Then, once the physician is provided with the set of recommended classes for each CPG she can select the most appropriate.

The interesting part of this approach is that while the physician manufactures the concept hierarchy she also inserts some prior knowledge on the desired organization of data. Actually, each new concept added to the hierarchy is usually labeled by a few keywords describing the supposed semantic meaning of its content. Moreover, the concept is related to other concepts (more specific, more general, related to, etc.). This prior knowledge is exploited by the

proposed model in order to perform a preliminary classification of CPGs according to their contents and the desired organization within the hierarchy.

The evaluation of the proposed approach has been performed on a set of real data selected from the above mentioned NGC database. The promising results showed that the approach can be valuable in order to create and populate new electronic hierarchical repositories of CPGs.

In Section 2 some research works related to medical knowledge management are discussed. Section 3 gives a description of the addressed task with some references to works aiming at solving similar problems. Section 4 introduces the model used to test the proposed solution. Section 5 describes the experimental setup. Finally, Sections 6 and 7 discuss the results of experiments and draw some conclusions respectively.

## 2 Related Works

Traditional text retrieval systems use the "vector space model" in which terms are extracted from the document and represented by either their term frequency as a *bag-of-words* or their term presence/absence as a *set-of-words*. The limitation of this approach is that humans search using concepts instead of terms. In the medical domain, concept-based search refers to a text retrieval approach where the documents are mapped to concepts based on their contents. SAPHIRE system [Hersh and Greens, 1989], for example, uses an approach in which concepts used for indexing are automatically extracted from the document. Actually, within biomedical domains, documents and queries are often mapped into a large vocabulary such as MeSH (see <http://www.nlm.nih.gov/mesh>) or UMLS [Humphreys and Lindberg, 1993], which is one of the major resources offered by the National Library of Medicine. The concepts in these vocabularies are represented in a hierarchical structure. This approach is somewhat limited, since users aren't always familiar, while querying, with the concepts in these vocabularies. Moreover, several studies had shown that such implementation of concept-based search might actually decrease the retrieval performance [Hersh and Hickam, 1993], mainly because there are no good automatic concept extractors.

This hierarchical organization of documents, also allows browsing through the concepts using the hierarchical structure. Such a browsing method forces the user to navigate the conceptual hierarchical structure. Alternatively, in these directories, searches can be limited to a specific concept and its sub-concept contents. However, in the medical domain documents are usually classified by a multitude of concepts, often as many as a dozen or even tens of concepts.

An example of solution to this problem is provided by Vaidurya, a concept based and context sensitive search engine for clinical guidelines [Moskovitch *et al.*, 2004]. This engine implements a concept based search where the user has to choose few concepts and the logic relation between them. In his query the user defines a relevant subset of the collection, based on the conceptual indexing.

Recent results had shown that searching within a hierarchical concepts indexing improved full text retrieval, even at the first and second level of the hierarchy, especially when using conjunctive queries [Moskovitch and Shahar, 2004]. However, in order to implement an accurate concept based search manual classification should be applied by an expert, a very time consuming task. Thus, an automatic hierarchical classifier for clinical guidelines is crucial. At least it can help during the manual classification recommending the most probable concepts to be assigned to the documents.

## 3 Task Definition

A concept hierarchy (also referred to as taxonomy) is a hierarchy of categories (also referred to as classes) which are represented as nodes in a tree. Each node is described in terms of both linguistic keywords (also referred to as labels) that ideally denote the "semantic meaning" of the nodes, and relationships with other categories. The leaves of the tree represent specific concepts, while nodes near the root of the tree represent more general concepts. In our particular task, each node of the hierarchy can contain CPGs and, in general, each CPG can belong to more than one category.

Annotation of document to classes is a typical task in information retrieval. The goal here is to identify the set of categories that best describe the content of an unclassified CPG. A wide range of statistical and machine learning techniques have been applied to text categorization (see for example [Ceci and Malerba, 2003; Chakrabarti *et al.*, 1997; Cheng *et al.*, 2001; Doan *et al.*, 2003; Dumais and Chen, 2000; Joachims, 1998; Jordan and Jacobs, 1994; Koller and Sahami, 1997; Ruiz and Srinivasan, 2002; Sun and Lim, 2001; Wang *et al.*, 1999; Weigend *et al.*, 1999]). However, none of the above models can be used to solve the proposed task. Actually, these techniques are all based on having some initial pre-labeled documents, which are used to train a (semi)-supervised model. Moreover, Although many real world classification systems have complex hierarchical structure, few learning methods capitalize on this structure. Most of the approaches above ignore the hierarchical structure and treat each category or class separately, thus in effect 'flattening' the hierarchical structure. In the case this hierarchical structure is kept the models only classify on the leaves of the structure.

These problems are partially solved by the way we use the *TaxSOM* model [Adami *et al.*, 2003b]. The model uses the prior knowledge to drive a clustering process and, as a result, it organizes the CPGs on a given concept hierarchy without any need of supervision during training. Basically, the model bootstraps the given taxonomy with a preliminary classification of CPGs that afterward need to be reviewed by the taxonomy editor.

The basic idea of the *bootstrapping* process is to support and alleviate the manual labeling of a set of unlabeled examples, providing the user with an automatically determined preliminary hypothesis of classification. The idea is to exploit the linguistic and the relational information encoded within a taxonomy through an unsupervised learning model. The paper illustrates how *TaxSOM* can be used to learn the prior knowledge encoded within a concept hierarchy in order to perform this preliminary classification of

CPGs.

In particular, the task goal is to provide the user with a list of recommended classes for each CPG, i.e., the most probable $k$ classes to which the CPG could belong.

## 4 Classification Models

A strategy to classify documents using prior knowledge is proposed by Yang [Yang, 1994]. Unlabeled documents are classified according to the lexical information associated to the categories. Specifically, a reference vector is built for each category, through the encoding of its labels. The documents are then associated to the category having the nearest reference vector (a standard prototype–based minimum error classifier). In the following, this simple class of keyword matching algorithms will be referred to as *baseline* categorization approach.

This classification method uses only lexical information, while topological information is neglected. To also use the hierarchical information we revised the *baseline* model according our scenario. Specifically, hierarchical knowledge was exploited building codebooks through the encoding of all labels in the current node and in all its ancestors, i.e., all labels of the nodes in the path from the root to the current node.

The above idea has been developed even more in the *TaxSOM* model [Adami *et al.*, 2003b]. Specifically, a *TaxSOM* is a collection of computational units connected so as to form a graph having the shape isomorphic to a given taxonomy. Such computational units, namely codebooks, are initialized as for *baseline*. Then an unsupervised training algorithm (similarly to Self Organizing Maps [Kohonen, 2001]) adapts these codebooks in order to take into account both the documents similarity and the constraints determined by the labels and the relationships. The basic idea is that once a *TaxSOM* has been properly trained the final configuration of the codebooks describes a clustered organization of documents that tailors the desired relationships between concepts.

The learning procedure of a *TaxSOM* is designed as an iterative process that can be divided into two main stages: a competitive step and a cooperative step. During *competitive* step the codebook most similar to the current input vector (a document) is chosen as the *winner* unit. In the *cooperative* stage all codebooks are moved closer to the input vector, with a learning rate proportional to the inverse of their topological distance from the winner unit. The iterations of the two steps are interleaved with an additional phase where the codebooks are constrained by the *a priori* lexical knowledge localized on the nodes.

## 5 Experimental Setup

We used the NGC CPGs collection to evaluate the suggested approach. The CPGs in the NGC hierarchy are classified along two hierarchical concept trees, Disorders and Therapies. Each concepts tree has roughly 1,000 unique concepts, in some regions the concepts trees are 10 levels deep, but the mean is 4 to 6 levels. There are 1136 CPGs, each CPG may have multiple classifications at different nodes by both concept trees and within the same tree. The classification is not necessary only on the leaves. CPGs have a mean of 10 classifications, while there exist CPGs classified by 90 concepts.

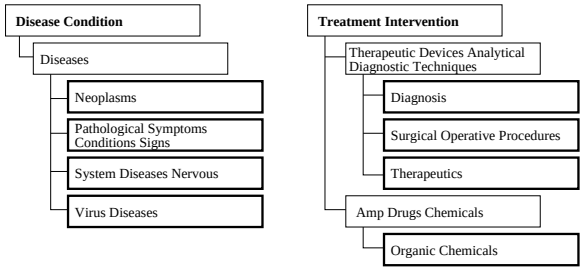


Figure 1: The eight selected taxonomies are subtaxonomies of the two original concept hierarchies. Specifically, the eight leaves in the above two trees (dark bordered boxes).

To evaluate the model with a plurality of datasets, we decided to split down the two original dataset ("treatment intervention" and "disease condition") into eight smaller and different datasets (see Figure 1). These datasets were selected according to dimensional criteria decided in the beginning of our testing process – their depth (i.e., how far the leaves are from their root), the number of nodes and the number of CPGs. The variability of both topics and dimensions allows the evaluation of the model without biases due to any prior knowledge, such as topic vocabulary, dimension of taxonomy, number of classes for each CPG, etc.

Table 1 summarizes the statistics of the selected taxonomies. It can be seen that the depth of the hierarchies ranges from 5 to 9 layers, with few hundreds nodes. While the number of CPGs range from hundreds to thousand. More interestingly, many nodes are not represented by any CPG, therefore, supervised classifiers cannot be learnt on these datasets. The main characteristic of such datasets is that usually the leaves are not empty, while the interior nodes (i.e., nodes that appear as parent of other nodes) many times are empty (sometimes more that $50\%$ of times).

Each taxonomy was preprocessed separately. The content of documents and the category labels were cleaned removing stop–words (articles, conjunctions, and prepositions) and reducing the vocabulary (i.e., the vector space representation) to 500 important keywords plus the labels of nodes. The important keywords were selected using the notion of Shannon Entropy[1]. Finally, CPG contents were encoded with a *set–of–words* representation (i.e., binary vectors).

As previously outlined, since to our knowledge there are not models devised to solve the proposed bootstrapping problem, we compared *TaxSOM* with the simple approach based on keyword matching refer to as *baseline*.

The model was tested on each taxonomy performing an hypothesis of classification for all CPGs, and the results were then compared with the original labeling. Actually, the addressed task requires the multi-classification of CPGs, therefore, given a CPG, both models generate a membership value for each class. These membership val-

[1] Shannon entropy is a standard information theoretic approach that can be used to measure the amount of information provided by the presence of a word in the dataset.

| *statistics* | *taxonomies* diagnosis | neoplasms | organic chem. | pathol. sympt. cond. signs | surgical operat. proced. | system diseases nervous | therap. | virus |
|---|---|---|---|---|---|---|---|---|
| max tree depth | 7 | 8 | 9 | 8 | 5 | 7 | 6 | 6 |
| tot nodes | 278 | 230 | 326 | 214 | 210 | 318 | 247 | 124 |
| tot docs | 1248 | 501 | 367 | 516 | 396 | 606 | 929 | 432 |
| min docs/node | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| max docs/node | 20 | 20 | 20 | 20 | 16 | 20 | 20 | 20 |
| average docs/node | 4.49 | 2.18 | 1.13 | 2.41 | 1.89 | 1.91 | 3.76 | 3.48 |
| min w/doc | 52 | 66 | 40 | 563 | 582 | 587 | 571 | 704 |
| max w/doc | 463 | 387 | 444 | 5132 | 5050 | 5726 | 6074 | 10599 |
| average w/doc | 218 | 223 | 232 | 1633 | 1378 | 1573 | 1707 | 261 |
| docs on leaves | 67% | 63% | 85% | 62% | 67% | 66% | 66% | 61% |
| % of leaves | 66% | 55% | 48% | 62% | 68% | 56% | 63% | 52% |
| empty nodes | 14% | 23% | 46% | 21% | 15% | 25% | 10% | 30% |
| empty leaves | 1% | 3% | 6% | 4% | 5% | 9% | 0% | 0% |
| empty interiors | 39% | 46% | 84% | 50% | 35% | 46% | 26% | 64% |

Table 1: Statistics of the selected concept hierarchies. The first group of rows describe the trees' dimension. The second group describes the datasets' dimension. The third group the documents' dimension. While the last two groups of rows describe respectively the distribution of CPGs in the hierarchies.

ues are then used to rank the classes, and this ranking is then used to select the best classes to recommend to the user. To evaluate the proposed two models we devised a specific measure – the *multi-classification k-coverage precision*. This measure allows a comparison of models rather than an objective evaluation.

The measure counts the percentage of CPGs "correctly" classified with respect to the total number of CPGs. The meaning of *k-coverage* is strongly related to the definition of "correct classification". In this case, a document is correctly classified when all the classes to which it belong are in the first $k$ recommended classes. The idea is that the system provide the user with a set of probable classes with which to label a give CPG. If all the interesting classes are among the recommended $k$ then the document is correctly classified.

For example, suppose we know that a CPG should be classified to three specific classes. If the model proposes all the three classes among the $k$ recommended, then the document is considered correctly classified. If, on the contrary, the model fail to propose at least one of the three classes in the recommended $k$ classes then the CPG is considered wrongly classified.

For example, a model having *k-coverage* equal to $60\%$ for $k = 10$ means that for $60\%$ of the documents all the corresponding correct classes appear in the first ten ranked classes.

## 6 Discussion of Results

The evaluation of the proposed model has been performed on all 8 smaller taxonomies and the two original taxonomies determining the *k-coverage* for all possible $k$s. In Figure 2 are depicted the graphs of the *k-coverage* for all $k$s for all the eight smaller taxonomies. It can be easily seen that for almost all reasonable $k$s the proposed *TaxSOM* model always outperform the *baseline* approach.

Actually, providing the best $k$ ranked classes for each CPG (where $k$ should be reasonable small to be explored by the physician) the probability of finding all the correct classes is higher for *TaxSOM* than for *baseline*. This means

| *taxonomies* | *k-coverage* *baseline* $k = 10$ | $k = 20$ | $k = 10\%$ | *TaxSOM* $k = 10$ | $k = 20$ | $k = 10\%$ |
|---|---|---|---|---|---|---|
| **diagnosis** | 11.3 | 23.3 | 30.9 (28) | 32.9 | 46.8 | 55.9 (28) |
| **neoplasms** | 38.2 | 46.2 | 48.1 (23) | 47.2 | 63.7 | 67.5 (23) |
| **organic chem.** | 60.7 | 71.0 | 79.3 (33) | 73.1 | 80.0 | 81.4 (33) |
| **pathol. sympt.** | 42.8 | 55.4 | 56.8 (21) | 64.2 | 76.5 | 77.9 (21) |
| **surgical op.** | 44.1 | 70.4 | 72.0 (21) | 68.8 | 76.9 | 78.0 (21) |
| **system dis.** | 26.1 | 38.9 | 50.0 (32) | 53.5 | 71.2 | 79.6 (32) |
| **therapeutics** | 23.9 | 39.2 | 40.9 (25) | 52.5 | 69.0 | 73.4 (25) |
| **virus diseases** | 27.5 | 48.9 | 30.5 (12) | 58.0 | 72.5 | 59.5 (12) |
| **disease cond.** | 8.0 | 14.0 | 55.3 (283) | 21.6 | 34.7 | 80.0 (283) |
| **treatment int.** | 3.9 | 5.9 | 38.3 (299) | 11.2 | 20.0 | 57.0 (299) |

Table 2: Results of *baseline* and *TaxSOM k-coverage* with three different values for $k$.

that the recommendations made by *TaxSOM* are "more correct" than those made by the *baseline* approach. Notice that the curves sometimes intersect with high values of $k$. In this case, however, the result is less interesting. In fact, the system is used to recommend the best classes, and the number of suggested classes should be as small as possible. Actually, in a real task, what we can expect from such a system is that for each CPG few classes are recommended as probable labeling classes for the given CPG.

In Table 2 are shown the results for three different situations: *(i)* a case where the system suggest a selection of 10 possible classes; *(ii)* a case where the system suggest a selection of 20 classes; *(iii)* a case where the system select the $10\%$ most probable classes among all classes. For all the three cases and for all taxonomies it is always valuable to use *TaxSOM* driven by the prior knowledge encoded in the taxonomy than just using *baseline* which only uses the keywords that best represent the concepts.

In the table we also provided the results of the same type of analysis done for the original two NGC hierarchies. From these results it can be seen that the behavior of the models is (obviously) influenced by the absolute number of classes in the hierarchy. Nonetheless, *TaxSOM* is still better that the *baseline* approach. Moreover, looking at the results of the third case (i.e. $k = 10\%$) the model still give interesting results also for the two big hierarchies.

## 7 Conclusions and Future Work

In the paper we presented an approach for helping physicians to organize CPGs into hierarchies of concepts. The challenge was twofold: to avoid the need for labeled documents in advance and to exploit relational knowledge encoded by a taxonomy. Experimental evaluation on a collection of CPGs gave the empirical evidence of the potential benefit for physicians while using the proposed model.

## References


[Adami *et al.*, 2003a] G. Adami, P. Avesani, and D. Sona. Bootstrapping for hierarchical document classification. In *Proc. of CIKM-03, 12th ACM Int. Conf. on Information and Knowledge Management*, pages 295–302. ACM Press, New York, US, 2003.

[Adami *et al.*, 2003b] G. Adami, P. Avesani, and D. Sona. Clustering documents in a web directory. In *Proc. of WIDM-03, 5th ACM Int. Workshop on Web Information and Data Management*, pages 66–73. ACM Press, New York, US, 2003.

[Ceci and Malerba, 2003] M. Ceci and D. Malerba. Hierarchical classification of html documents with webclassii. In *Proc. of the 25th European Conf. on Information Retrieval (ECIR'03)*, volume 2633 of *Lecture Notes in Computer Science*, pages 57–72, 2003.

[Chakrabarti *et al.*, 1997] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Using taxonomy, discriminants, and signatures for navigating in text databases. In M. Jarke, M.J. Carey, K.R. Dittrich, F.H. Lochovsky, P. Loucopoulos, and M. A. Jeusfeld, editors, *VLDB'97, Proc. of 23rd Int. Conf. on Very Large Data Bases*, pages 446–455. Morgan Kaufmann, 1997.

[Cheng *et al.*, 2001] C.H. Cheng, J. Tang, A.W.C. Fu, and I. King. Hierarchical classification of documents with error control. In *PAKDD 2001 - Proc. of 5th Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, volume 2035 of *Lecture Notes in Computer Science*, pages 433–443, 2001.

[Doan *et al.*, 2003] H. Doan, P. Domingos, and A. Halevy. Learning to match the schemas of data sources: A multistrategy approach. *Machine Learning*, 50:279–301, 2003.

[Dumais and Chen, 2000] S. Dumais and H. Chen. Hierarchical classification of web document. In *Proc. of the 23rd ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR'00)*, 2000.

[Grimshaw and Russel, 1993] J.M. Grimshaw and I.T. Russel. Effect of clinical guidelines on medical practice: A systematic review of rigorous evaluations. *Lancet*, pages 1317–1322, 1993.

[Hersh and Greens, 1989] W.R. Hersh and R.A. Greens. Saphire - an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval and hierarchical relationships. *Computers and Biomedical Research*, 23:410–25, 1989.

[Hersh and Hickam, 1993] W.R. Hersh and D.H. Hickam. A comparison of two methods for indexing and retrieval from a full text medical database. *Medical Decision Making*, 13(3):220–26, 1993.

[Humphreys and Lindberg, 1993] B.L. Humphreys and D.A. Lindberg. The umls project: making the conceptual connection between users and the information they need. *Bull Med Libr Assoc*, 81(2):170–177, 1993.

[Joachims, 1998] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning (ECML '98)*, 1998.

[Jordan and Jacobs, 1994] M.I. Jordan and R.A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6:181–214, 1994.

[Kohonen, 2001] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Series in Information Sciences*. Springer, Berlin, 2001.

[Koller and Sahami, 1997] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. In D.H. Fisher, editor, *ICML 1997, Proc of the 14th Int. Conf. on Machine Learning*, pages 170–178. Morgan Kaufmann, 1997.

[McCallum and Nigam, 1999] A. McCallum and K. Nigam. Text classification by bootstrapping with keywords, em and shrinkage. In *In ACL Workshop for Unsupervised Learning in NLP*, 1999.

[Moskovitch and Shahar, 2004] R. Moskovitch and Y. Shahar. Effective concept-search in hierarchical organized library. Technical Report ISE-TR-314/2004, Dept. of Information Systems Engineering, Ben Gurion University, 2004.

[Moskovitch *et al.*, 2004] R. Moskovitch, A. Hessing, and Y. Shahar. Vaidurya - a concept-based, context-sensitive search engine for clinical guidelines. In *Proc. of the joint conf. of AMIA04 and Medinfo-2004*, San Francisco, CA, US, 2004.

[Ruiz and Srinivasan, 2002] M.E. Ruiz and P. Srinivasan. Hierarchical text categorization using neural networks. *Information Retrieval*, 5(1):87–118, 2002.

[Sun and Lim, 2001] A. Sun and E.P. Lim. Hierarchical text classification and evaluation. In N. Cercone, T.Y. Lin, and X. Wu, editors, *ICDM 2001 - Proc. of the 2001 IEEE Int. Conf. on Data Mining*, pages 521–528. IEEE Computer Society, 2001.

[Wang *et al.*, 1999] K. Wang, S. Zhou, and S.C. Liew. Building hierarchical classifiers using class proximity. In *Proc. of the 25th VLDB Conference*, 1999.

[Weigend *et al.*, 1999] A.S. Weigend, E.D. Wiener, and J.O. Pedersen. Exploiting hierarchy in text categorization. *Information Retrieval*, pages 1(3) 193–216, 1999.

[Yang, 1994] Y. Yang. Expert network: effective and efficient learning from human decisions in text categorization and retrieval. In *Proc. of the 17th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 13–22, 1994.

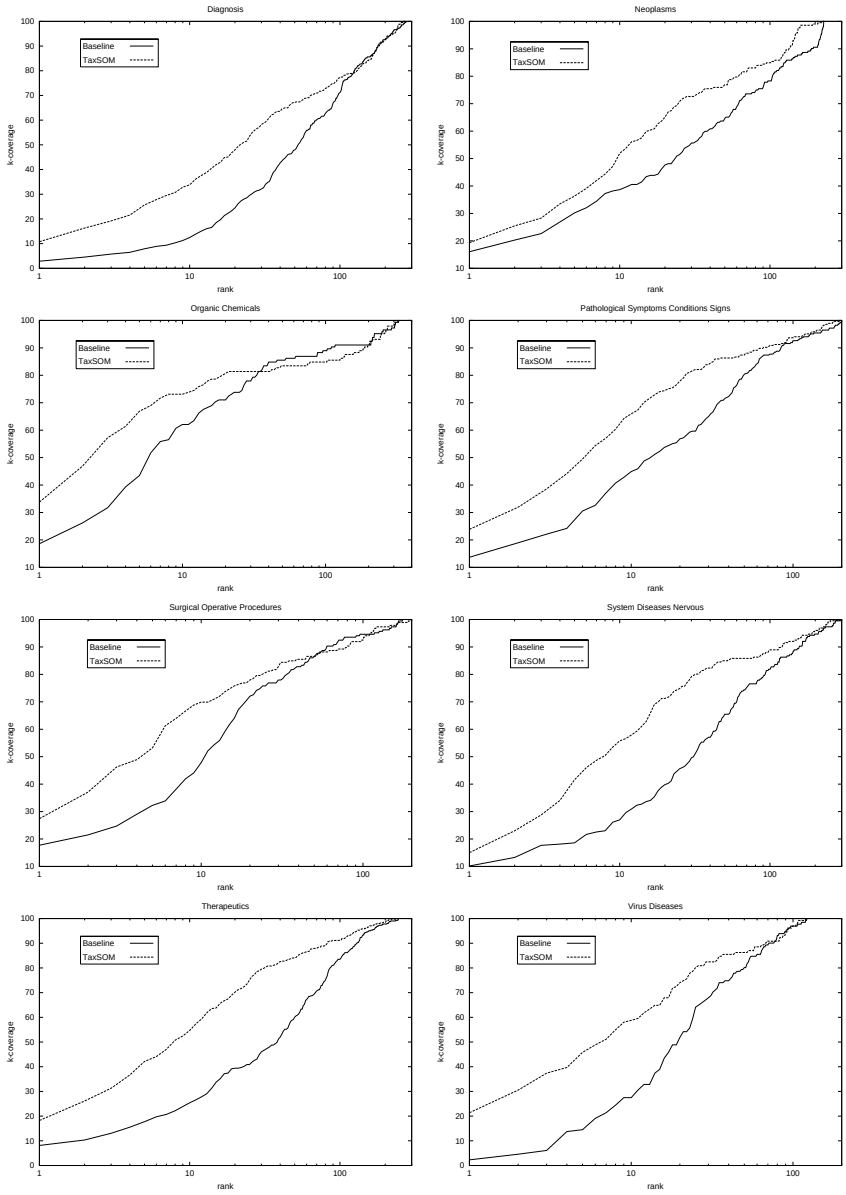


Figure 2: The graphs depict the *k-coverage precision* for all eight datasets.