RJ10365 (A0510-017) October 28, 2005
Computer Science

IBM Research Report

On the Use of Hierachy Information in Mapping Patents to Biomedical Ontologies

Luo Si, Tapas Kanungo
IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

ON THE USE OF HIERARCHY INFORMATION IN MAPPING PATENTS TO BIOMEDICAL ONTOLOGIES

LUO SI[†] AND TAPAS KANUNGO

IBM Almaden Research Center, 650 Harry Road,
San Jose, CA 95120, USA

Abstract

Ontologies provide a vocabulary to reason about classes of entities. In biomedical research numerous ontologies like MeSH [28], SNOMED [25], and Gene Ontology [24] are being created and used for analyzing and reasoning about diseases, symptoms, root causes on one end, and discovering associations by mapping one ontology to another, on the other. However, research on mapping ontologies has been limited to mapping ontologies based on research documents. In this paper we argue that the ontology representing patents is very valuable since it represents the intellectual property owned by businesses and also is a means of associating chemical and biological entities that impact human health to most other manufactured goods and manufacturing processes. In this context we create a tool that allows a user to map patents into the MeSH ontology and find related topics and discover relationships that would otherwise not been possible if one looks at patents in isolation. We create human ground truth mappings that represent the "correct" answers for both training our proposed learning algorithms, and comparing the algorithm generated mappings to the human-generated ones. We show that the performance of the mapping algorithm improves substantially when we incorporate the structure information of MeSH and Patent ontologies.

[†] This work was done when Luo Si was visiting IBM Almaden Research Center.

1. Introduction

Biomedical information overload is a blessing in disguise. Without the right tools one can be buried in it. However, the right data mining tools and data can allow us to discover relationships, find new cures, and prevent possible health risks. In this article we propose that patents, which record intellectual property related to biomedical inventions (such as drugs), can add value to various biomedical data such as Medline [30], Medical Subject Headings (MeSH) [28], Unified Medical Language System (UMLS) [25] and Gene Ontology (GO) [24] that are currently used for biomedical information search. To illustrate this point, consider the following scenarios:

Latent relationships. A company invents and patents a new drug for a specific disease. Another company finds that the critical function accomplished by the drug can also be used to fix another health problem. However, the relationship is discovered by finding the relationship amongst concepts using an ontology such as MeSH, but not the original patent ontology. How can we make such discoveries more automated?

Prior art search. A patent attorney wants to know what prior art exists on a specific patent application. Besides finding Medline documents and patents by key words, the examiner would like to be able to find the associated MeSH categories, and extend the scope of the search to semantically related concepts and co-occurring concepts using MRREL (semantically related concepts) [25] and MRCOC (co-occurrences between concepts) [25]. Currently the examiner would have to do this manually.

Health hazards. Let's say that a biomedical document mentions that a specific chemical can have an adverse effect on the health of an individual, and goes on to specify specific symptoms that one might have if the chemical is consumed or inhaled. An unrelated patent document describes the use of the same chemical in a manufacturing process of a component or a toy. With the current information tools, it is not possible to discover this fact automatically [7].

Tech transfer. Let's say a researcher in academia invents a new drug and publishes a paper on the topic. Next the researcher would like to do more serious tests to verify the viability of the drug. One way to get funded for the work would be to look at companies that work on the topic, and related concepts in the patent corpus. Similarly a company might want to find academic researchers working on related concepts. Currently we would have to do this manually by searching documents within Medline and Patent datasets and relating them via the MeSH ontology one-by-one.

In each scenario listed above, we notice that an automatic association discovery tool would be beneficial. While one can still discover the associations today with existing tools, it is very labor intensive, time consuming, and prohibitively expensive. In this paper we describe a tool that is a step towards this automation process. We use the structure information in the patent and MeSH ontologies to create a classification algorithm that maps a patent document to semantically closest concepts in MeSH. As a by-product the tool also generates a list of most similar documents in the Medline and patent corpus.

We propose several algorithms to map patent documents into MeSH ontology. Simple K-nearest neighbor (KNN) [21] algorithms that do not utilize ontology structure information are first described in this paper and serve as baseline algorithms. More complicated algorithms are proposed to take advantage of structure information among source International Patent Classification (IPC) [29] patent ontology and target MeSH ontology. In addition, a finer category-specific mapping algorithm is proposed to better justify the mapping pattern with respect to different MeSH ontology categories. Empirical studies show that the category-specific mapping algorithm that also utilizes ontology structure information results in more than 20 percent improvement for different evaluation measures over simple KNN algorithm.

2. Literature Survey

The work described in this article builds on research conducted on ontology mapping and text categorization for biomedical and patent documents.

Ontology mapping related work has been aggressively pursued in the biomedical, semantic web, computational linguistics, and the database communities. A good overview on different types of ontology mapping problems and approaches is described in Noy [13], including database schema mapping techniques tackled in the database community. In some cases, ontology mapping depends only on the descriptions of the categories to create a mapping. In other cases, including the work described in this article, the mapping takes advantage of the "instances" available within each category. Noy and Musen [14] describe an interactive tool that helps domain experts to align various ontologies. The ultimate goal of Cantor *et al.* [3] is to discover relationships between diseases and genes by learning the mapping between Systematized Nomenclature of Human and Veterinary Medicine (SNOMED) [25] disease ontology and the gene ontology using the MeSH ontology as an intermediary. Our work, on the other hand, tries to build a relationship between biomedical documents and patent documents by using the structure information provided by MeSH and Patent ontologies. While there are a few TREC-like biomedical ontology competitions starting up [32,33], training corpus for instance-based ontology mapping does not exist. This forced us to create our own groundtruth dataset.

Ontology mapping has been a quite active research topic in semantic web community (Berners-Lee, Hendler, and Lassila [2]). Agrawal and Srikant [1] propose a probabilistic method of mapping web taxonomies like that of Google and Yahoo using common web documents as instances for training and for validation. Doan *et al.* [5] use a relational matching approach to create a mapping between ontology nodes using instances. They apply their technique to web taxonomy mapping and department course taxonomy mapping. Zhang and Lee [22] use various statistical techniques like boosting to integrate multiple web taxonomies. Finally, Resnik [15] uses an information-theoretic approach to assess semantic similarity between taxonomies for linguistic disambiguation.

Another related topic is text analysis and categorization, which deals with building statistical classifiers using words and other features from individual documents. The classification work has been applied to various types of text documents including biomedical documents and patent documents. However the purpose of this result was

Table 1. The distribution of identified top level MeSH categories of 253 patent documents (total 1,148).

Assigned MeSH Top Category	**Counts**	**Assigned MeSH Top Category**	**Counts**
A: Anatomy	184	H: Physical Sciences	11
B: Organisms	13	I: Anthropology, Education, Sociology and Social Phenomena	1
C: Diseases	90	J: Technology and Food and Beverages	20
D: Chemicals and Drugs	188	L: Information Science	14
E: Analytical, Diagnostic and Therapeutic Techniques and Equipment	453	M: Persons	8
F: Psychiatry and Psychology	13	N: Health Care	55
G: Biological Sciences	98		

not to build bridges across document corpuses, but to classify documents within its type (e.g. Medline or patent corpus) or learn a new hierarchy based on some principles. Chakrabarti *et al.* [4] and Dumais and Chen [6] learn hierarchical taxonomies from web documents. Yang [21] quantitatively compares various statistical text clustering algorithms using the Ohsumed corpus, and Kiritchenko, Matwin, and Famili [9] learn hierarchical text classifiers for associating genes with gene ontology codes. A review of various statistical clustering techniques is provided by Jain, Murty, and Flynn [8]. Larkey [11] and Koster, Seutter, and Benney [10] have applied text categorization techniques to patent documents where they try to automatically associate a patent with its correct category.

3. Data Description

The main data sets that we work with are Medline [30] and Patents [27] databases, and MeSH [28] and IPC [29] ontologies. The MeSH ontology is organized into a tree structure with 15 top level categories such as A (anatomy), B (organisms) etc, while each of them is in turn divided into many subcategories. The current MeSH ontology contains 42,611 nodes and has up to 12 levels. In order to provide a more general MeSH structure for patent documents, and also to assure better quality of ground truth assignment by human judges, the MeSH tree is pruned to the third level (e.g., Heart:

A07.541). This pruning procedure leaves 1,514 categories. A subset of the Ohsumed [31] data is utilized to learn the content representation of MeSH ontology. This subset contains 133,749 documents published from 1987 to 1989 in the Medline database.

The patent documents studied in this paper come from the United States Patent and Trademark Office (USPTO) [27] database. All the patent documents in USPTO database are assigned with category/categories from the International Patent Classification (IPC) ontology. The IPC ontology has a depth of 5 levels and contains 8 top level categories (e.g., A: Human Necessities). Patent documents from three IPC categories (A61 -- medical or veterinary science, hygiene; C07 -- organic chemistry and C12 -- biochemistry) are selected for our biomedical mapping work. This selection procedure was validated by checking the patent distributions of large biomedical companies such as Pfizer and Novartis. In particular, patent documents in these categories approved in 1990 and 1991 are used.

Two human judges provide ground truth data of MeSH categories for 253 patent documents. First, a simple K near-nearest neighbor algorithm is applied on these patent documents to generate MeSH category candidates for each patent document; then the judgment of each candidate in top 20 categories for each patent document is provided by the human judges. This procedure produces 1,148 identified categories for the 253 patent documents. The detailed distribution of top level MeSH categories for these patent documents is shown in Table 1. The 253 patent documents were assigned to 13 out of 15 top level MeSH categories, all categories except K (Humanities) and Z (Geographic Locations).

4. Problem Statement

The patent documents are originally organized in the source IPC ontology and our task is to map them into the target MeSH ontology. In this section we formally define the mapping problem.

Each patent document $\vec{d}$ is represented by a set of word $\{w_1, w_2, \ldots, w_l\}$and a set of attribute features showing its position in the source ontology. We denote the source ontology as *S* and the target ontology as *T*. Both the source and target ontologies contain hierarchical information. For example, target ontology *T* includes a set of categories as $\{c_1, c_2, \ldots, c_t\}$ while all the category nodes are organized in a hierarchical tree (each node has a single direct ancestor while there is a single root for the whole tree). The Ancestor(c) operator obtains the direct ancestor of a particular category node c. Category node c belongs to another category node c' as long as c' is the (direct or indirect) ancestor of category c; this is denoted as $c \in c'$. The structure of source ontology *S* is defined in a similar way as target ontology *T*.

Given all the available information for target ontology *T* and source ontology *S*, the ontology mapping task is to identify whether a particular patent document $\vec{d}$ should be assigned into a particular category node c in the MeSH ontology. To put this decision procedure into probabilistic setting, we calculate the probability that the patent document $\vec{d}$ should be put into MeSH category c as $P(A=1|\{\vec{d}, c, S, T\})$. A is a binary random variable which is 1 when the document should be assigned into the category and 0 otherwise.

5. Ontology Mapping Algorithms

We now describe simple K-nearest-neighbor-based ontology mapping algorithms and several more sophisticated algorithms that utilize structure information of ontologies.

5.1 *K-Nearest neighbor based ontology mapping*

One straightforward algorithm to map patent documents into MeSH ontology is to ignore the structure information in both IPC patent document ontology and MeSH ontology and to apply standard text categorization technique (e.g. K-nearest neighbor [21]) for flattened MeSH categories (all the category nodes are treated separately).

In the testing phase, each patent document $\vec{d}$ is compared against each Medline document $\vec{d}_{med}$ in the Ohsumed corpus using a similarity measure. Our similarity measure is derived from Okapi information retrieval system:

$$Sim(\vec{d},\vec{d}_{med}) = \sum_{w \in (\vec{d}_{med} \wedge \vec{d})} belief(w,\vec{d},\vec{d}_{med}) \qquad (1)$$

which calculates the accumulated belief scores for all overlap words that appear in both the patent document and the Medline document. The word belief score is calculated according to the Okapi formula [26].

After the similarity scores of the patent document and Medline documents are calculated, the K top-ranking Medline documents are identified as $KNN(\vec{d})$ and are used to calculate the score for each MeSH category as:

$$Score(c,\vec{d}) = \sum_{c \in \vec{d}_{med} \wedge \vec{d}_{med} \in KNN(\vec{d})} Supp(\vec{d},\vec{d}_{med}) \qquad (2)$$

where $Supp(\vec{d},\vec{d}_{med})$ is the evidence contributed by document $\vec{d}_{med}$ for classifying $\vec{d}$ as category c. The equally weighted KNN algorithm treats all categories in the set of K nearest neighbors equally by setting $Supp(\vec{d},\vec{d}_{med}) = 1$; while the score weighted KNN algorithm favors categories from Medline documents with higher similarity scores as $Supp(\vec{d},\vec{d}_{med}) = Sim(\vec{d},\vec{d}_{med})$. Finally, all the MeSH category candidates can be sorted with respect to these scores. If binary decisions have to be made, some thresholding strategies should be applied [21].

5.2 *Utilizing hierarchical structure information of ontologies for mapping*

One limitation of the simple K-nearest neighbor algorithm is that it does not utilize the structure information of either the source IPC ontology or the target MeSH ontology. Categories within local structure of source and target ontologies (e.g., category nodes with the same direct ancestor category node) generally tend to represent similar semantic concepts, and also categories within source ontology may share similar semantic concepts with corresponding category nodes within target ontology. This type of local semantic consistency and cross ontology semantic relationship may provide useful information for the mapping of patent documents into target MeSH ontology.

For example, for a specific MeSH category candidate for a patent document, if many other top-ranking category candidates are found to be in a small local structure of the category, it is intuitive to assign a higher probability of choosing this category because of more semantic supporting evidence from other similar categories. This example shows the advantage of utilizing the structure information of the MeSH target ontology to improve the mapping accuracy. In particular, three features are formally defined to represent the target ontology information. First, a group of L (i.e., 20 in this work) top-ranking category candidates for document $\vec{d}$ is identified as $LTopCat(\vec{d})$ by a simple K-nearest neighbor algorithm. Then, for a specific category candidate c, additional supporting evidence from similar categories is defined as:

$$f_{t\arg 1}(c,\vec{d},T) = \sum_{c'\in Ancestor(c)\wedge LTopCat(\vec{d})} Score(c',\vec{d}) \qquad (3)$$

The first feature of MeSH target ontology sums up the supporting evidence of all category nodes that belong to the director ancestor of the category node in consideration. The second feature considers a larger local structure, which includes all top-ranking category nodes that belong to the second level ancestor of the category node in consideration as:

$$f_{t\arg 2}(c,\vec{d},T) = \sum_{c'\in Ancestor(Ancestor(c))\wedge c'\in LTopCat(\vec{d})} Score(c',\vec{d}) \qquad (4)$$

Finally, an additional normalization feature is introduced, which sums up the supporting evidence of all top-ranking categories:

$$f_{t\arg 3}(\vec{d},T) = \sum_{c'\in LTopCat(\vec{d})} Score(c',\vec{d}) \qquad (5)$$

Equation 3, 4 and 5 introduce additional evidence to the direct evidence between category c and patent document $\vec{d}$, which is denoted as the feature $f_{direct}(\vec{d},T) = Score(c,\vec{d})$. Then the next question is how should we combine the direct and indirect evidence. What weights should be associated with this set of evidence when they vote for category candidates. We accomplish this evidence integration by learning the weights of an exponential model using training data. Formally, the probability of assigning a particular category c in the MeSH target ontology to patent document $\vec{d}$ is calculated as:

$$P_{targ}(A=1|\{\vec{d},c,S,T\})=\frac{\exp(\lambda_1 * f_{direct}+\lambda_2 * f_{targ1}+\lambda_3 * f_{targ2}+\lambda_3 * f_{targ3}+\lambda_4 * f_1)}{1+\exp(\lambda_1 * f_{direct}+\lambda_2 * f_{targ1}+\lambda_3 * f_{targ2}+\lambda_3 * f_{targ3}+\lambda_4 * f_1)} \tag{6}$$

This model is denoted as M_{targ} and maps ontologies using the structure information of target ontology. The weights $\lambda_j\{1 \le j \le 4\}$ are associated with different features. The bias feature f_1 is always set to 1. These weights are estimated using the training data as follows. Let O be the number of training documents. The training goal is to maximize the log-likelihood of correctly assigned category nodes of those documents:

$$\vec{\lambda}^*_{targ} = \underset{\vec{\lambda}}{\arg\max} \sum_{i=1}^{O} \sum_{c \in LTopCat(\vec{d}_i)} \log P_{targ}(A = a_{true} | \{\vec{d}_i, c, S, T\}) \tag{7}$$

where $\vec{\lambda}^*_{targ}$ is the set of estimated feature weights of M_{targ}, a_{true} is the true annotation (whether a category belongs to a document or not). This is a convex optimization problem and many algorithms have been proposed for it. We use the Quasi-Newton algorithm [12] to estimate the model parameters as it has been demonstrated to be better than several other alternatives. With the estimated weight parameters, we can rerank the category candidates from the results of simple K-nearest neighbor algorithm for better accuracy.

It is reasonable to assume that the structure information of source ontology is also valuable. For example, for a specific MeSH category candidate of a patent document, if many other patent documents in the same IPC category of this specific patent document also rank the category highly, it is more probable that the category is a correct assignment of the patent document. Based on this intuition, four features are introduced in this work to utilize the structure information of source IPC ontology:

$$f_{src1}(c,\vec{d},S) = \frac{1}{C\{\vec{d}' \in Ancestor(\vec{d})\}} \sum_{\vec{d}' \in Ancestor(\vec{d})} \frac{Score(c,\vec{d}')}{\sum_{c' \in LTopCat(\vec{d}')} Score(c',\vec{d}')} \quad (8)$$

This feature averages all the (normalized) supporting evidence of category c from patent documents that belong to the same ancestor of the patent document in consideration. The averaging and normalization procedures are introduced to consider the variation of number of patent documents in local source ontology structure and also the variance of the supporting evidence score of patent document and target ontology category (i.e., normalize the document length factor as introduced in Equation 1). $C\{\vec{d}' \in Ancestor(\vec{d})\}$ denotes the number of patent documents that belong to the same ancestor of the patent document in consideration. The second feature f_{src2} averages the supporting evidence of all category nodes that belong to the direct ancestor of the category node in consideration from the set of patent documents in a similar way as the first feature. Two more features f_{src3} and f_{src4} sum up supporting evidence in a similar way as f_{src1} and f_{src2} but from a set of patent documents that belong to the second level ancestor of the patent document in consideration.

Furthermore, an exponential model associated with direct evidence f_{direct} and the four features that utilize structure information of source IPC ontology can be defined similarly as Equation 6. This model is denoted by M_{src} and creates mapping by using the structure information of the source ontology or M_{src}. It is also trained with the Quasi-Newton method.

Finally, it is straightforward to utilize the structure information of both the target MeSH ontology and the source IPC ontology. In particular, we built an exponential model that incorporates the f_{direct} feature associated with the direct evidence, the 3 features associated with the target ontology information, the 4 features associated with source ontology information and 1 constant bias feature. We denote this model as M_{targ_src}.

5.3 *Category-Specific mapping algorithm*

In Section 5.2 we presented several learning algorithms that utilize structure information of target ontology and source ontology to improve the accuracy of mapping patent documents into MeSH ontology. One particular issue about these algorithms is that they apply the same model to calculate the probabilities of choosing different categories for a particular document. This general model may be too coarse for distinguishing different categories. For example, if all the features have the same supporting evidence for different features, the particular document will have the same probabilities to be assigned into different categories. This may not be correct if there are some categories that appear much more frequently than other categories.

Therefore, a finer category-specific algorithm is proposed to build different models for different types of categories. As there are more than 1,500 categories in the pruned 3 level MeSH ontology, in order to avoid the overfitting problem of limited amount of training data, only 13 top level (i.e., shown in Table 1) category-specific models are built. All the training data that belong to a top level category is used to build one model. In order to alleviate the data sparseness problem for some small top level categories (e.g., 2 top level categories have less than 10 positive training examples), a Laplacian prior is associated with the parameters of category-specific model. Specifically, a general model (e.g. M$_{targ}$ in Equation 6) is first built with all training examples. This set of model parameters is used as prior for the category-specific model. For example, a category-specific model for top level category "E" based on M$_{targ}$ can be described as:

$$P_{E\_t\arg}(A=1|\{\vec{d},c,S,T\})=\frac{\exp(\lambda_{E\_1}*f_{direct}+\lambda_{E\_2}*f_{t\arg1}+\lambda_{E\_3}*f_{t\arg2}+\lambda_{E\_3}*f_{t\arg3}+\lambda_{E\_4}*f_{1})}{1+\exp(\lambda_{E\_1}*f_{direct}+\lambda_{E\_2}*f_{t\arg1}+\lambda_{E\_3}*f_{t\arg2}+\lambda_{E\_3}*f_{t\arg3}+\lambda_{E\_4}*f_{1})} \qquad (9)$$

and is trained with the maximum a posterior criterion as:

$$\vec{\lambda}^{*}_{E\_t\arg}=\arg\max_{\vec{\lambda}}(\sum_{i=1}^{O}\sum_{c\in LTopCat(\vec{d}_i)}\log P_{E\_t\arg}(A=a_{true}|\{\vec{d},c,S,T\})-r\left\|\vec{\lambda}-\vec{\lambda}_0\right\|_1) \qquad (10)$$

where $\left\|\vec{\lambda}-\vec{\lambda}_0\right\|_1$ represents the Laplacian prior with respect to the general model $\vec{\lambda}_0$ and r reflects the weight of prior, which is arbitrarily set to 10 in the work. Note that this model is trained only for MeSH categories under a specific top level category (e.g., E).

Category-specific models based on M_{src} and$_{targ\_src}$ can be built similarly.

6. Experimental Protocol

One natural measure to evaluate the effectiveness of mapping patent documents into MeSH ontology is to use the *F1* measured used for text categorization. Specifically, *F1* measure is derived from the measure of precision and recall. Precision (*P*) is the proportion of correctly chosen category nodes among all the chosen category nodes. Recall (*R*) is the proportion of correctly chosen category notes by a mapping algorithm among all the correct category nodes. *F1* is the harmonic mean of precision and recall: *F1=2PR/(P+R).*

However, one implicit requirement of using Recall measure is that all correct category nodes for every testing patent document have been identified. This may not be true as human judges only provide ground truth data for top 20 categories ranked by simple KNN algorithm of each patent document. This set of correct categories may not be complete. For example, if some MeSH categories do not appear in Medline documents of the Ohsumed subset, they cannot be labeled by human judges. Therefore, in addition to the F1 evaluation measure, we report a rank-based Precision measure. This measure evaluates the precision of correctly identified category nodes in the top ranking categories by averaging across all testing patent documents.

In the empirical study, 100 out of 253 annotated patent documents are randomly selected and used for training while the rest are used as testing data. For a more thorough analysis, the random split process is repeated ten times for each set of experiment and the evaluation results are averaged. Although simple KNN algorithm does not utilize the training data, its effectiveness is still measured on the same set of 10 random split data as other algorithms for the comparison.

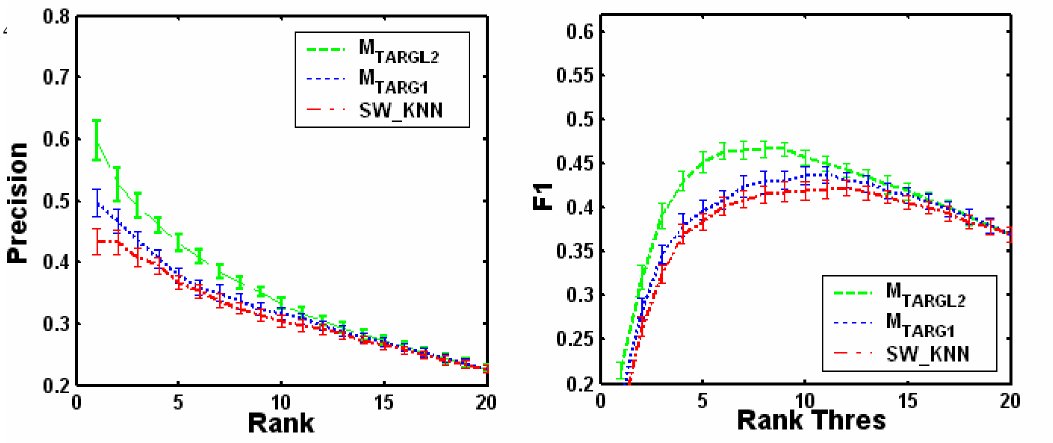

Fig 1 (a)

Fig 1 (b)

Fig 1 (a). Averaged precision of top-ranking categories for testing patent documents by mapping algorithm using structure information of two level target ontology (M_{TARGL2}), or one level target ontology (M_{TARGL1}), and the baseline score weighted KNN (SW_KNN) algorithm. Fig 1 (b). Averaged F1 measure with respect to different rank cut thresholds by these algorithms.

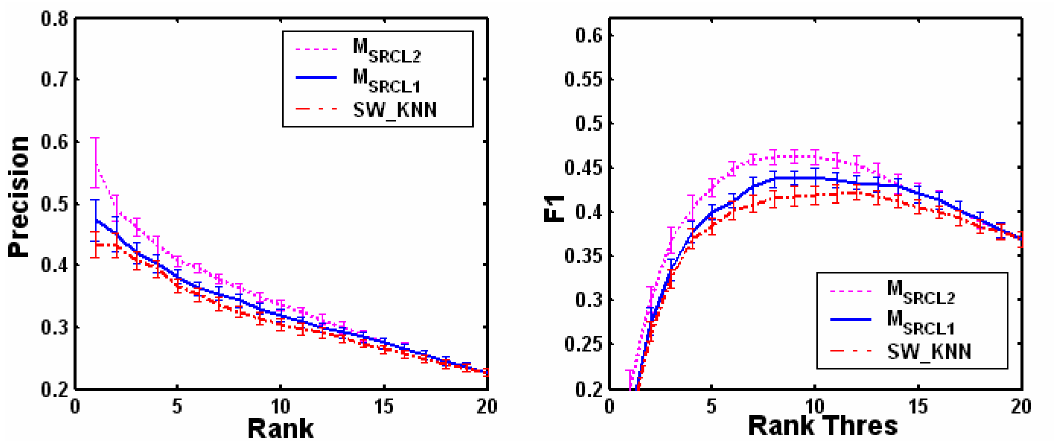

Fig 2 (a)

Fig 2 (b)

Fig 2 (a). Averaged precision of top-ranking categories for all testing patent documents by mapping algorithm using structure information of two level source ontology (M_{SRCL2}), or one level source ontology (M_{SRCL1}), and the baseline score weighted KNN (SW_KNN) algorithm. Fig 2 (b). Averaged F1 measure with respect to different rank cut thresholds by these algorithms.

7. Experimental Results and Discussion

We now first discuss the empirical results of simple KNN algorithms; then we show the effectiveness of mapping algorithms by utilizing ontology hierarchical structure

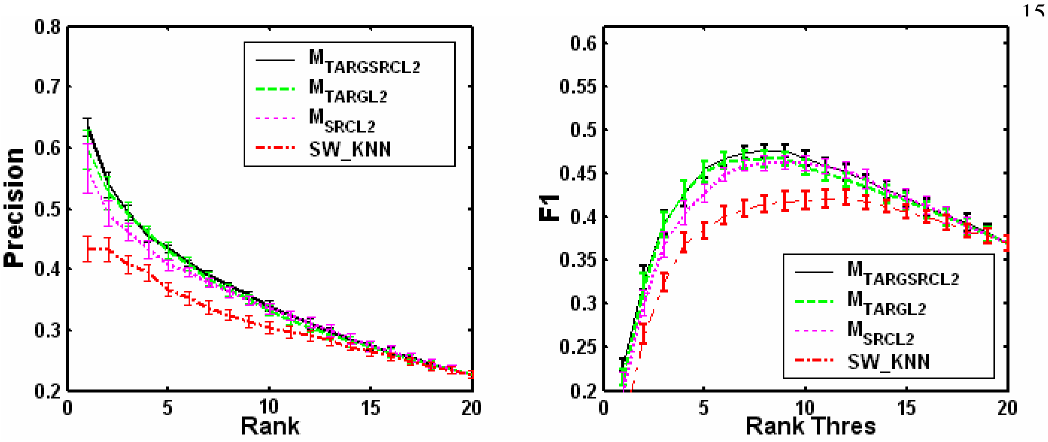

Fig 3 (a) Fig 3 (b)

Fig 3 (a). Averaged precision of top-ranking categories for all testing patent documents by mapping algorithm using structure information of two level target ontology (M_{TARGL2}), or source (M_{SRCL2}) ontology, or both target and source ontology ($M_{TARGSRCL2}$), and the baseline score weighted KNN (SW_KNN) algorithm. Fig 3 (b). Averaged F1 measure with respect to different rank cut thresholds by these algorithms.

information; and finally we demonstrate the advantage of category-specific mapping algorithm.

7.1 *Effectiveness of mapping algorithms that utilize hierarchical structure information of ontologies*

We conducted baseline empirical experiments with two simple KNN algorithms -- equally weighted KNN algorithm and score weighted KNN algorithm -- which are described in Section 5.1. Our experiments show that the two simple KNN algorithms have very similar performance. Therefore, only the result for score weighted KNN algorithm is reported.

The next set of experiments demonstrates the power of mapping algorithms that utilize ontology structure information. Two models that use structure information of target ontology are compared with the score weighted KNN baseline algorithm in Figure 1. One model M_{TARGL1} uses one-level structure information as features f_{targ1} and f_{targ3} (i.e., normalization feature), while the other model M_{TARGL2} utilizes all the two level features f_{targ1}, f_{targ2} and $f_{targ3.}$ It can be seen that the improvement of M_{TARGL1} over simple KNN algorithm is rather small while the advantage of method M_{TARGL2} is much larger. This demonstrates the importance of using more complex hierarchical ontology structure than the single level structure. The same set of experiments has been conducted to study the models that utilize structure information of source ontology. The results are shown in Figure 2 demonstrate the advantage of utilizing complex hierarchical ontology structure.

Another set of experiments show the performance of the $M_{TARGSRCL2}$ model. This model takes advantage of the two level structure information of both the target ontology and the source ontology. The results are shown in Figure 3. It substantially outperforms the simple KNN algorithm. However, it has very similar performance as that of $M_{TARGL2.}$ One possible explanation is that structure information of target ontology may provide more semantic information than the source ontology as the task is to map document into target ontology. This is more probably true in the case of ontology structures with fine granularity (e.g., 1,514 possible categories in our application).

7.2 *Effectiveness of category-specific mapping algorithms*

In Section 5.3 we discussed the potential problem of using the same general model like M_{targ_src} for all category candidates and have proposed a category-specific mapping algorithm. The power of this finer algorithm is seen in the experimental results shown in Figure 4. This figure compares the results of category-specific $M_{TARGSRCL2SPEC}$ model and the original $M_{TARGSRCL2}$ model that uses a single general model for all categories. It can be seen from Figure 4 that category-specific $M_{TARGSRCL2SPEC}$ model has a large advantage over the original general $M_{TARGSRCL2}$. This suggests the importance of finer category-specific model. More experiments have been conducted to study the category-specific M_{targ} and M_{src} models, and both of them show improvement over their original versions.

Finally, the large advantage of category-specific $M_{TARGSRCL2SPEC}$ model over the baseline score-weighted KNN algorithm explicitly shows the advantage of building effective ontology mapping algorithm. A more detailed analysis shows that the category-specific $M_{TARGSRCL2SPEC}$ algorithm achieves 24 percent improvement on

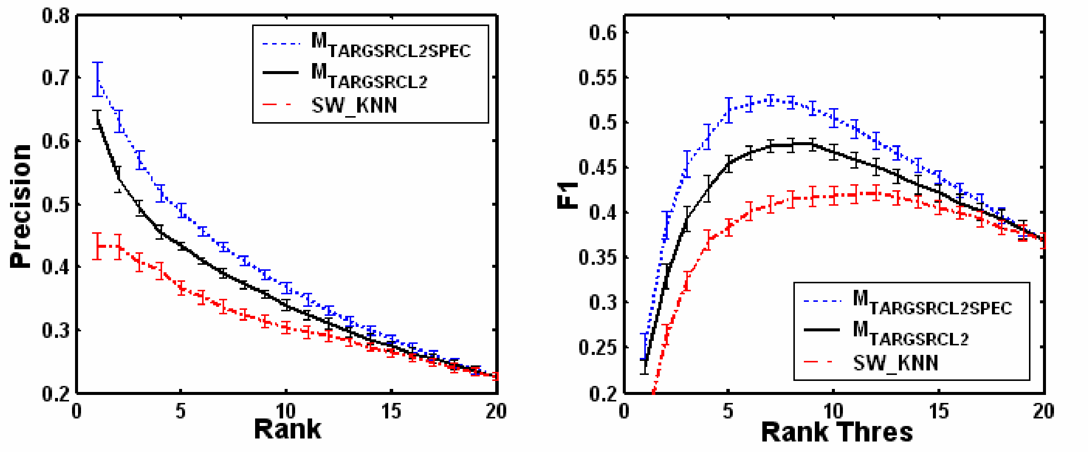

Fig 4 (a) Fig 4 (b)

Fig 4 (a). Averaged precision of top-ranking categories for testing patent documents by category-specific $M_{TARGSRCL2SPEC}$ algorithm, the original $M_{TARGSRCL2}$ algorithm and score weighted KNN algorithm. Fig 4 (b). Averaged F1 measure with respect to different rank cut thresholds by these algorithms.

average. This is calculated by computing the percentage improvement of average of ranked-based precision values against the average of simple KNN algorithm (see Figure 4(a)). The improvement of maximum F1 measure is 24.4 percent (see Figure 4(b)).

8. Discussion

In our work we have used only the hierarchy information in the MeSH and IPC labelings. When the ontologies have richer structural information (e.g. DAGs), the computational methodology outlined in our paper remains the same – extract local features and estimate the weights corresponding to the features using the training dataset. The algorithm is quite dependent on the features and we have proposed no systematic way of coming up with an "optimal" set of features. These features have to be hypothesized and then selected using the standard pattern recognition methodologies using training and test data.

Ontologies are manually curated and represent a rich amount of knowledge. The success of the ontology mapping algorithm heavily depends on the manually curated "mappings" between the ontologies. While we have used manually created mappings to compare our algorithm-generated mappings, one has to realize that mappings created by two humans are not going to agree most of the time. In fact, in the information retrieval community, it has been found that “inter-rater agreement” between two human curators is less than 50% [17,18]. Furthermore it has been reported [19,20] that if human judgments are “pooled,” prior to use in algorithm training, the performance of the algorithm is in general improves. Thus, for ontology mapping, too, the performance of the algorithm will depend on the manually created ontology mappings, which is a laborious process. However, bootstrapping approaches and online relevance feedback methods can speed up the process dramatically and keep improving the mapping while users use the system.

9. Conclusion

Automatic mapping of patents to MeSH ontology opens up new ways to drug discovery, detecting possible health hazards, intellectual property search, etc. We presented a text analysis tool that helps us move closer to this objective of automated discovery of associations. We presented a learning algorithm that was trained using human judgments. We showed that we get about 24% better mapping performance when we use the structure information present in the two ontologies. In this work we used ontologies that are simple structured thesauri. It will be interesting to apply our techniques on more complex ontologies (e.g., ontologies within UMLS [25]). We intend to extend this work by using chemical and gene annotations, cross-references, and other sources like the SNOMED [16,25] and OMIM [23] corpuses.

Acknowledgements

We will like to thank Steve Boyer, Jeffrey Kreulen, James Rhodes, Scott Spangler, and David Martin for many useful discussions.

References

1. R. Agrawal and R. Srikant, “On Integrating Catalogs,” in *Proc. of the 10th Int. World Wide Web Conference (WWW)*, 2001.
2. T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web,” *Scientific American*, 284(5), pp. 34-43, 2001.
3. M. N. Cantor, I. N. Sarkar, O. Bodenreider, and Y. A. Lussier, “GENESTRACE: Phenomic Knowledge Discovery Via Structured Terminology,” in *Proc. of the Pacific Symp. on Biocomputing*, 2005.
4. S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan, “Using Taxonomy, Discriminants, and Signatures for Navigating in Text Databases,” in *Proc. of Int. Conf. on Very Large Databases*, Athens, Greece, pp. 446–455, 1997.
5. A. Doan, J. Madhavan, R. Dhamankar, P. Domingos, A. Halevy, “Learning to match Ontologies on the Semantic Web,” *VLDB Journal,*12,pp.303-319, 2003.
6. S. Dumais and H. Chen, “Hierarchical Classification of Web Content,” in *Proc. ACM SIGIR Conf.,* Athens, Greece, pp. 256-263, 2000.
7. S. P. Gardner, “Ontologies and semantic data integration,” *Drug Discovery Today*, vol. 10, pp. 1001-1007, July 2005.
8. A. K. Jain, M. N. Murty, and P. J. Flynn, “Data Clustering: A Review,” *ACM Computing Surveys,* 31(3), pp. 264-323, 1999.
9. S. Kiritchenko, S. Matwin, and A. F. Famili, “Hierarchical Text Categorization as a Tool of Associating Genes with Gene Ontology Codes,” in *Proc. Work. on Data Mining and Text Mining in Bioinformatics*, pp.30-34, 2000.
10. C. H. A. Koster, M Seutter, and J. Beney, “Multi-classification of Patent Applications with Winnow,” in *Proc. Ershov Memorial Conf.*, pp.546-555, 2003.
11. L. S. Larkey, “A patent search and classification system,” in *Proc. of the ACM Conf. on Digital Libraries,* Berkeley, CA, pp. 79-87, 1999.
12. T. Minka, “A comparison of numerical optimizers for logistic regression,” unpublished draft.

13. N. F. Noy, “Semantic Integration: A Survey of Ontology-Based Approaches,” *SIGMOD Record,* 33(4), pp. 65-70, 2004.
14. N. F. Noy and M.A. Musen, “The PROMPT suite: Interactive tools for ontology merging and mapping,” *Int. J. of Human-Computer Studies*, 59, pp.983–1024, 2003.
15. P. Resnik, “Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language,” *Journal of Artificial Intelligence Research,* vol. 11, pp. 95-130, 1999.
16. K. A. Spackman, “Normal forms for description logic expressions of clinical concepts in SNOMED RT,” in Proc AMIA Symp, pp. 627-631, 2001.
17. E. M. Voorhees, “Variations in relevance judgements and the measurement of retrieval effectiveness,” in Proc. of the SIGIR Conf. on Research and Development in Information Retrieval, Melbourne, Australia, pp. 315-323, 1998.
18. E. M. Voorhees, “The philosophy of information retrieval evaluation,” in Proc. Workshop of the Cross Language Evaluation Forum, Portland, OR, 20002.
19. W. J. Wilbur, “A comparison of group and individual performance among subject experts and untrained workers at the document retrieval task,” *Journal of the American Society for Information Science,* 49, pp. 517-529, 1998.
20. W. J Wilbur, “The knowledge in multiple human relevance judgements,” *ACM Transactions on Information Science,* vol. 64, pp. 155-169, 1999.
21. Y. Yang, “An evaluation of statistical approaches to text categorization,” *Information Retrieval,* 1, pp. 69-90, 1999.
22. D. Zhang and W. S. Lee, “Learning to Integrate Web Taxonomies,” *Journal o f Web Semantics,* 2(2), pp. 131-151, 2004.
23. OMIM, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
24. Gene Ontology, http://www.geneontology.org/
25. UMLS, http://www.nlm.nih.gov/research/umls
26. Lemur, http://www.lemurproject.org/
27. USPTO, http://www.uspto.gov
28. MeSH, http://www.nlm.nih.gov/cgi/request.meshdata

29. IPC, http://www.wipo.int/classifications/en/ipc

30. MEDLINE, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed

31. OHSUMED, ftp://medir.ohsu.edu/pub/ohsumed

32. http://co4.inrialpes.fr/align/Contest/

33. http://www.atl.external.lmco.com/projects/ontology/i3con.html