Disease Similarity Calculation on Simplified Disease Knowledge Base for Clinical Decision Support Systems

Mai Omura
Nara Institute of Science & Technology
8916-5 Takayama, Ikoma,
Nara 630-0192, Japan

Yuka Tateishi
National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo 101-8430, Japan

Takashi Okumura
National Institute of Public Health,
2-3-6 Minami, Wako-shi,
Saitama 351-0197, Japan

Abstract

For clinical decision support systems designed to help physicians make diagnostic decisions, "disease similarity" data is highly valuable in that they allow continuous recommendation of diagnostic candidates. To build such a recommendation algorithm, this paper explores a method to measure disease similarity between diseases on a simplified disease knowledge base. Our disease knowledge base comprises disease master data, symptom master data, and disease–symptom relations that include clinical information of 1550 disorders. The calculation of the disease similarity is performed on this knowledge base, with i) standardized disease classification, ii) probabilistic calculation, and iii) machine learning, and the results are evaluated with a gold standard list audited by a physician. We also propose a novel metric for evaluation of the algorithms to calculate the disease similarity. A comparative study between the algorithms revealed that the machine learning approach outperforms the others. The results suggest that even a superficial calculation on a simplified knowledge base may satisfy the clinical needs in this problem domain.

Introduction

Clinical Decision Support Systems (CDSSs) are systems that assist physicians to make diagnostic and therapeutical decisions (Berner 2007). For most implementations in this class of systems, they are programmed to present disease candidates upon input of clinical findings of patients by physicians. In such a candidate list, CDSSs may provide a list of similar diseases for a disease that a user physician focuses, so that the system can realize continuous recommendation of diagnostic candidates.

The calculation of disease similarity lies within the core of such a recommendation system, which measures the similarity of any given diseases. The simplest approach here is to count the number of overlapping findings that the diseases may present. However, clinical manifestations of diseases have a variety of modalities, and the similarity is not easily calculated. For example, a disease may present a high body temperature, which might be coded by a code for *fever*. Nevertheless, there exist various variants of fever, such as *high fever*, *low grade fever*, *continuous fever*, and *intermittent fever*. Similarly, a simple symptom, *abdominal pain*, may have modifiers such as *stabbing abdominal pain*, and *abdominal pain that goes to right lower quadrant of abdomen*.

These examples suggest that symptoms of diseases are not simply expressed as a vector of symptoms. On the other hand, it is very expensive to express exact knowledge of existing diseases, which may number in the several thousands, and thus, a fully-fledged knowledge base is not a reasonable choice to take for most occasions. Accordingly, the best strategy here is to find a way to calculate a similarity metric of reasonable performance on a simple vector-oriented disease knowledge base.

In this paper, we tried to calculate disease similarly on such a simplified disease knowledge base and evaluated several possible approaches. First, the next section outlines the materials used: the simplified disease knowledge base and the training data. The section is followed by the description of the proposed methods for the similarity calculation. Then, the systematic evaluation is provided, coupled with the detailed discussion. We then present related work, and conclude the paper in the last section.

Disease Knowledge Base and Similarity Data

The simplified disease knowledge base encodes the knowledge of diseases into a very simple structure (Okumura et al. 2014). This section descries the disease knowledge base, which comprises disease master data, symptom master data, and disease-symptom relations, coupled with the similarity data used for the evaluation of the proposed approaches.

Disease Master Data

The disease master data is a table in our disease knowledge base that contains basic attributes of the diseases in the knowledge base, such as "Alzheimer's disease" (D_{72}), "Gastric cancer" (D_{534}), and "Allergic rhinitis" (D_{835}). The disease master data contains information of 1550 diseases, and defines their unique IDs, names, synonyms, prevalence and standardized disease codes. The prevalence of a disease is simplified into a four-scale measure, as shown in Table 1. The value of prevalence for a disease is referred to as $P(D_i)$ throughout this paper. Note that the disease master data does not have relation data between diseases.

Table 1: Prevalence scale of a disease

Scale	Frequency $P(D_i)$	Remark
4	0.01	Common disease
3	0.001	Ordinary disease
2	0.00001	Rare disease
1	0.000001	Very rare disease

Table 2: Frequency scale of a symptom

Scale	Frequency $P(S_i)$	Remark
4	0.055	Common symptom
3	0.0055	Occasional symptom
2	0.000055	Rare symptom
1	0.0000055	Very rare symptom

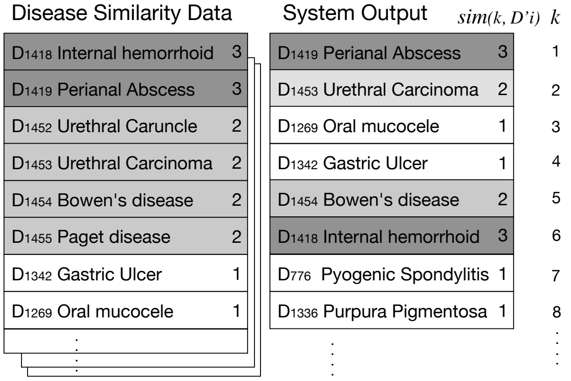
Each disease in this knowledge base has disease codes, expressed in the International Classification of Diseases (ICD). ICD is a systematic classification of diseases built by the World Health Organization (World Health Organization 2011), and covers a vast class of diseases. Because ICD is a systematic classification, it is highly likely that diseases with similar ICD codes have similar properties. However, the ICD system is made in an arbitrary way, and the number of disease items in each level of the ICD hierarchy is unbalanced. Further, clinical manifestations of two diseases with similar ICD codes may not necessarily look alike. Accordingly, the ICD codes are not a universal foundation for similarity calculation of diseases. Toward appropriate calculation of disease similarity, we need other metrics, in addition to the classification of the diseases in question, such as prevalence and symptoms.

Symptom Master Data

The symptom master data is also a table in our disease knowledge base for clinical findings. The data includes 597 symptoms, such as “Fever” (S_{260}), “Jaundice” (S_{394}), “Abdominal pain” (S_{808}). The symptom master data also carry frequency of symptoms, in a four-scale measure, as shown in Table 2. The value for a symptom is referred to as $P(S_i)$ throughout this paper.

Although the symptoms are listed in a uniform way, they differ in various aspects and their importance is not identical. For example, even a healthy person may occasionally experience malaise (S_{256}), but altered level of consciousness (S_{127}) strongly suggests a serious disorder happening in the patient. For this reason, each symptom is given a parameter, *significance*, in a five-scale measure. The significance score expresses medical importance of each symptom, implying higher seriousness as the number steps up. The significance score of a symptom S_i is expressed as $sig(S_i)$, in the later discussions.

The prevalence of diseases and the frequency of symptoms are substantial indexes that are statistically measurable. However, in reality, it is not an easy task to fix the frequency of all the diseases and the symptoms comprehensively and precisely. Accordingly, we choose to use subjective estimation of a physician, the last author of this article, for the frequency parameters to calculate disease similarity in this paper.



Disease Similarity Data		System Output	sim(k, D'i)	k
D_{1418} Internal hemorrhoid	3	D_{1419} Perianal Abscess	3	1
D_{1419} Perianal Abscess	3	D_{1453} Urethral Carcinoma	2	2
D_{1452} Urethral Caruncle	2	D_{1269} Oral mucocele	1	3
D_{1453} Urethral Carcinoma	2	D_{1342} Gastric Ulcer	1	4
D_{1454} Bowen's disease	2	D_{1454} Bowen's disease	2	5
D_{1455} Paget disease	2	D_{1418} Internal hemorrhoid	3	6
D_{1342} Gastric Ulcer	1	D_{776} Pyogenic Spondylitis	1	7
D_{1269} Oral mucocele	1	D_{1336} Purpura Pigmentosa	1	8
⋮		⋮		⋮

Figure 1: An example of disease similarity data and system output.

Disease-Symptom Relation

The disease-symptom relation is a data set that expresses relations between diseases and symptoms, in the above-noted masters. In our simplified knowledge base, each relation has a subjective probability of a disease, given the existence of the symptom in a patient. For example, “Jaundice” in a patient may suggest *Hepatitis* and *Hemolytic anemia*. Accordingly, we have two relations “Jaundice → Hepatitis” and “Jaundice → Hemolytic anemia”, each of which has a subjective probability in the knowledge base. They are, in fact, diagnostic contribution of each symptom, and referred to as $P(D_i|S_j)$ throughout the paper.

Gold-Standard and Evaluation Metric

For evaluation of the disease similarity we propose, it is desirable that an objective and quantitative metric for the similarity is available, coupled with the gold-standard data. However, it is difficult to define the similarity between diseases in a quantitative manner, and subjective probability is used in the disease knowledge base for simplification, as mentioned. Accordingly, for evaluation of the proposing calculations, we decided to perform comparison of the rankings in the list of similar diseases that the calculations generate, with the list for similar diseases, not the absolute scores. To this end, we built such lists for 80 common diseases in the disease master data, subjectively evaluated by a physician, against the 1550 diseases.

During the compilation process, we noticed that a physician cannot define strict ordering of diseases in the similarity metric, although there are rough order between them. For example, Influenza is more relevant to Allergic rhinitis, than Appendicitis, but it is hardly possible for physicians to determine which is more relevant to rhinitis, Influenza or Common cold. They perceive all the differences qualitatively, not quantitatively, and the ordering in the list for similar diseases can reflect rough positions in the physicians mind. Thus, we decided to classify the diseases in the disease master data, roughly into three classes, according to the similarity to a disease in question: similar (3), related (2), and unrelated (1). The left side of Figure 1 shows an example of disease similarity data for “External Hemorrhoid” (D_{871}).

For evaluation of the algorithms, there must be a method to compare the output of the algorithms, which is a list of diseases sorted by the magnitude of similarity, with the handmade gold-standard in the three-scale measure. For this comparison, we attempted to use a metric, Normalized Discounted Cumulative Gain (NDCG) (Järvelin and Kekäläinen 2000). NDCG is a popular measure for ranking quality, which evaluates not only binary relevance but graded relevance scale ranking, as well. This measure first defines DCG (Discounted Cumulative Gain), as follows, which is the sum of the scores for the topmost p diseases in a ranking list (Burges et al. 2005).

$$DCG_p(\mathbb{D}'_i) = \sum_{k=1}^{p} \frac{2^{sim(k,\mathbb{D}'_i)} - 1}{\log_2(k+1)} \tag{1}$$

$\mathbb{D}'_i$ is defined as system's output for D_i, as an ordered list whose elements are similarity of diseases. For example, the right-side instance of Figure 1 is defined as $\mathbb{D}'_i = (3, 2, 1, 1, 2, 3, 1, 1, ...)$. In this equation, $sim(k, \mathbb{D}'_i)$ is k-th disease in the disease similarity data for D_i. Based on this measure, NDCG is defined as follows. $\mathbb{D}_i$ expresses the disease similarity data for D_i. In the case of Figure 1, the disease similarity data becomes $\mathbb{D}_i = (3, 3, 2, 2, 2, 2, 1, 1, ...)$.

$$NDCG_p(\mathbb{D}'_i) = \frac{DCG_p(\mathbb{D}'_i)}{DCG_p(\mathbb{D}_i)} \tag{2}$$

The normalized DCG is a reasonable metric to evaluate algorithms that generate ranking as output. However, our gold-standard is expressed in a three-scale measure, not a ranking, although the output of algorithms is made in a full-scale ranking. To bridge the gap, we need to cancel out the influence caused by ordering of diseases in the same similarity class. To this end, we define Normalized Disease Similarity Measure (NDSM), modifying the NDCG, as follows.

$$NDSM_p(\mathbb{D}'_i) = \frac{DSM_p(\mathbb{D}'_i)}{DSM_p(\mathbb{D}_i)} \tag{3}$$

In this metric, $DSM_p(\mathbb{D}'_i)$ and $f(x, k)$, used in $DSM_p(\mathbb{D}'_i)$, are defined as (4) and (5), respectively.

$$DSM_p(\mathbb{D}'_i) = \sum_{k=1}^{p} f(sim(k, \mathbb{D}'_i), k) \tag{4}$$

$$f(x,k) = \begin{cases} (len(\mathbb{D}_i) - k)/len(3) & (x = 3) \\ (len(\mathbb{D}_i) - k)/(len(3) + \\ \qquad\qquad\qquad len(2)) & (x = 2) \\ 0 & (x = 1) \end{cases} \tag{5}$$

where $len(\mathbb{D}_i)$ is the size of disease similarity data for D_i, and $len(x)$ is the number of items in a similarity class x for D_i. For example, the right-side list of Figure 1, generated by an algorithm, is evaluated as follows: $DSM_p(\mathbb{D}'_i) = (len(\mathbb{D}_{871}) - 1)/len(3) + (len(\mathbb{D}_{871}) - 2)/(len(3) + len(2)) + 0 + 0 + (len(\mathbb{D}_{871}) - 5)/(len(3) + len(2))...$

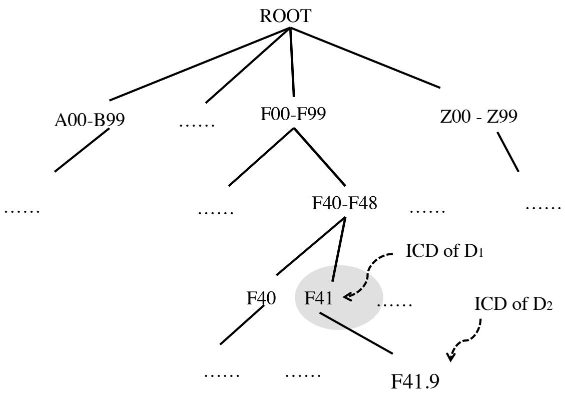


Figure 2: An example of ICD hierarchy

Owing to $f(x, k)$ function, $NDSM_p$ cancels out ordering in a same similarity class and prioritizes the high-similarity classes, contrary to NDG_p. Hereafter, parameter p is set to 250, because diseases below this line will not be helpful for user physicians of target CDSSs.

Disease Similarity Calculations

This section presents approaches to calculate the disease similarity we investigate in this study, namely, i) ICD-based disease similarity, ii) probability-based disease similarity, and iii) a machine learning-based approach.

1) ICD-based Disease Similarity (ICD)

It is natural to assume that diseases with similar ICDs are similar, as mentioned in the previous section, and thus, we attempted to quantify the distance between ICD codes. The ICD classification system has a hierarchical organization, exemplified in Figure 2. For diseases D_i and D_j, let d_i, d_j be the depth of ICD for disease d_i, d_j, respectively. Likewise, let $LCS(D_i, D_j)$ be the least common superconcept of ICDs for D_i, D_j. Now, the similarity between D_i and D_j is defined, as follows (Wu and Palmer 1994).

$$ICD_{dist}(D_i, D_j) = \frac{2 \times LCS(D_i, D_j)}{d_i + d_j} \tag{6}$$

In the case of Figure 2, disease D_1 has ICD “F41” and D_2 has “F41.9”. The superconcept of two ICDs is “F41”, and is calculated as $2 \times 3/(3.0 + 4.0) = 0.857$. Note that a disease may have two or more ICD codes. Accordingly, we extend $ICD_{dist}(D_i, D_j)$ as the least distance among the ICD pairs given. During the evaluation of approaches, we refer to the ICD-based similarity calculation as `ICD`.

ICD is a widely-used standard for disease classification. Consequently, $ICD_{dist}(D_i, D_j)$ can universally quantify the disease similarity between diseases. However, the relationship between diseases is not appropriately determined solely with the ICD distance. In fact, our preliminary experiment failed to detect similar diseases for “menopausal disorder”, “atopic dermatitis”, and so on. In such cases, it is necessary to discover relations of diseases, based on the clinical manifestations they present.

2) Probability-Based Disease Similarity (Prob)

Similarity between diseases can be evaluated in respect to their mechanism, as well as to their presenting symptoms. For evaluation of the similarity of diseases with a finite set of independent symptoms, the simplest approach is to calculate the Jaccard Coefficient (Jaccard 1912) between the set of symptoms for the diseases. However, even if the two diseases share identical symptoms, a symptom may occur more frequently in one disease, and the other may rarely present the symptom. Obviously, it is an oversimplification that easily leads to diagnostic errors. Accordingly, we attempted to take the frequency of symptoms into consideration here.

The calculation of symptomatic frequency requires the conditional probability $P(S_j|D_i)$, for any given set of diseases and symptoms. However, the disease knowledge base contains only $P(D_i|S_j)$, as descried in the previous section. Accordingly, we make rough Bayesian estimation of $P(S_j|D_i)$, utilizing probabilities $P(D_i|S_j)$,$P(S_j)$, and $P(D_i)$, stored in the knowledge base, as (7). Note that this calculation is inaccurate in nature, because all the probabilities are subjective.

$$P(S_j|D_i) = \frac{P(D_i|S_j)P(S_j)}{P(D_i)} \tag{7}$$

Additionally, we attempt to incorporate the significance of symptoms for the calculation of the probability-based similarity, because physicians would prioritize the similarity between symptoms of high medical importance, over trivial symptoms.

The calculation first assumes disease D_i and D_j as sets of symptoms, and extracts $|D_i \cap D_j|$ as S. Then, let S_x be the subset of the symptoms S with significance x. Likewise, let X be the set of value significance can take, which is $X = \{1, 2, 3, 4, 5\}$. Now, we define the probability-based similarity, as (8), incorporating the probability by the Bayesian estimation and the significance metric we introduced.

$$Prob(D_i, D_j) = \sum_{x \in X} x \log_{10} \sum_{s \in S_x} (1 + P(s|D_i)^2) \tag{8}$$

In (8), a logarithm is taken to avoid the case where matches in significant symptoms are overwhelmed by abundant symptoms of low significance. Also in (8), $P(s|D_i)$ is squared to emphasize the high-probability over the low-probability. The resulting approach, to calculate the similarity based on the probability, is referred to as Prob in the comparative study below.

3) A Machine Learning Approach (CRR)

The ICD-based approach and the probability-based approach have distinct characteristics, both of which have their own rationale. Accordingly, a deliberate hybridization of the approaches might outperform the two, exploiting the advantages and compensating for the shortcomings. For the combination of the approaches, we investigated machine learning techniques, and there are algorithms *to learn ranking of instances*, which have been used for many applications in Information Retrieval, Natural Language Processing, Data Mining, and others (Li 2011).

Table 3: Feature list for the machine learning approach, (CRR). The D_i and D_j are diseases, S is the set of symptoms shared by D_i and D_j and s is symptom $s \in S$.

Each symptom $s \in S$
The length of S
The conditional frequency of symptom $P(s|D_i)$
The conditional frequency of disease $P(D_j|s)$
The conditional frequency of symptom $P(s|D_i)$
The conditional frequency of disease $P(D_j|s)$
The significance of symptom $sig(s)$
The similarity of D_i and D_i by ICD $ICD_{dist}(D_i, D_j)$

For the calculation of disease similarity, the input query is a target disease D_i, and the input instance is the list of diseases excluding D_i. The expected output is the ranking list sorted by the disease similarity for D_i, trained by the gold-standard data. For example, in the case of Figure 1, the input query is “External Hemorrhoid” (D_{871}), the input instance is the list of diseases excluding D_{871}, and the output would be the ranking list of similar diseases for D_{871}.

There exists various models to learn the ranking of instances, and we selected Combined Regression and Ranking (CRR) (Sculley 2010). CRR is a method which combines pairwise rank-based and regression objectives using standard stochastic gradient. Sculley (2010) shows that CRR often achieves performance equivalent to the best of both ranking-only and regression-only approaches. The feature list used in our study is shown in Table 3. This setting incorporates the ICD-based and the probability-based approaches, and may reflect implicit knowledge of physicians that is encoded in the training data. We call this calculation method for the similarity of diseases CRR.

Experimental Results and Discussion

Experimental Setting

For the calculations of disease similarity, ICD and Prob can perform the calculation without any training. In contrast, CRR needs the disease similarity data to build the model. Accordingly, in the evaluation of CRR, the disease similarity data is used for leave-one-out cross-validation, using 79 diseases as the training data, and one disease for the evaluation of the output. For construction of the CRR model, the sofia-ml package [1] is used.

Results

Table 4 shows the experimental results of the proposed methods. Comparing the averages of the scores for 80 diseases each for the calculations, CRR outperformed the other two, and ICD performed better than Prob. The number of the best cases, in Table 4, indicates the number of diseases where each method scored the best, suggesting that

[1] https://code.google.com/p/sofia-ml/

Table 4: The results of disease similarity calculation. Each score is an average for the 80 diseases.

Method	$NDCG$	$NDSM$	# of best cases
ICD	0.846	0.707	16
Prob	0.815	0.734	5
CRR	**0.890**	**0.899**	59
Total	–	–	80

CRR is not always the best approach. For statistical significance, Wilcoxon signed rank test on the $NDSM$ scores confirmed CRR over ICD ($p = 0.001$), and CRR over Prob ($p = 0.001$), but the difference between Prob and ICD was not significant ($p = 0.20$).

There are 16 cases where ICD performed the best, which includes cases for "Deep-Seated Candidiasis" (D_{177}) and "IIa-type hyperlipoproteinemia" (D_{482}). Likewise, Prob outperformed the others in 5 cases, including "Allergic conjunctivitis" (D_{1380}) and "Hemorrhoid" (D_{1418}).

Analysis

The experiments clarified that CRR achieved the highest performance, on average. However, there are cases in which the others performed better. Figure 3 shows actual outputs for such an example, "(Deep-Seated) Candidiasis" (D_{177}). The figure extracts the topmost seven items that CRR and ICD algorithms selected as the most relevant.

Candidiasis is an infection of fungus into a variety of body parts, mostly skin and respiratory system at first, and occasionally occurs in patients whose immunity is compromised. This nature promotes CRR and Prob to favor diseases with dermatological and respiratory symptoms. However, because there are so many diseases that present such common symptoms, the chances might increase for these methods to inadequately choose such diseases.

Meanwhile, ICD is a classification system, based on etiology of diseases, and thus, the algorithm extracts diseases with similar mechanisms, in this case, diseases that are caused by fungus and observed in immunocompromised patients. Indeed, physicians would also favor such diseases, if they are asked to choose relevant diseases for Candidiasis, because they are characterized as such. Presumably, CRR might have lost cases for such a reason.

The difference between $NDCG$ and $NDSM$ indicates that the advantage of CRR becomes more remarkable, if the evaluation ignores the changes of rankings in a same class and prioritize the diseases in higher classes. This observation, as well as the advantage of CRR exemplified in Table 4, would hold true even for different knowledge bases, or even if we revise our data, considering the volume of the data we used.

Discussion

ICD picks up similar diseases based on the mechanism of the disease. Meanwhile, there are a variety of cases in which other types of similarity is more valuable to help physicians make diagnosis. For example, in the task to find similar diseases for certain diseases that are characterized by *abdominal pain* or *headache*, physicians would focus on the symptoms, not the mechanisms. There are various types of similarities and physicians switch their strategies to take, based on the contexts. The high performance of CRR is ascribed to the adapting nature of machine learning techniques.

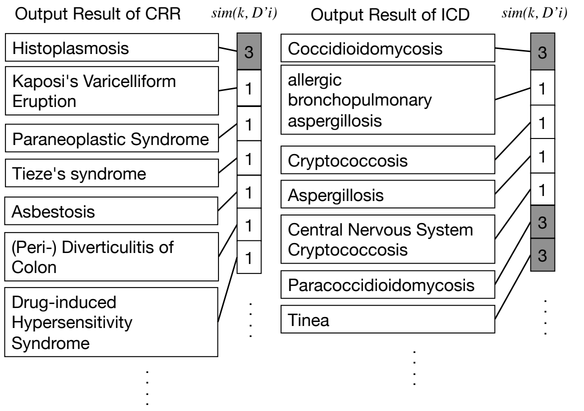


Figure 3: The example of CRR and ICD results of "(Deep-Seated) Candidiasis" (D_{177}) (top 7). The left side is the result of CRR, the right side is the result of ICD.

However, the experimental results exhibited 21 cases where ICD and Prob outperformed CRR, suggesting room for further improvement. The straightforward approach here is to find a factor that can predict the method of the highest performance, given a certain disease. In search for such a predictor, several hypotheses are possible. First and foremost, Prob would be advantageous if the symptoms includes significant ones. Second, a disease with only trivial symptoms might favor ICD. Third, the category of a disease, which is readily available in the ICD code, might predict the superior approach. Fourth, the significance of the symptom that appears most frequently in a disease might be the predictor. As illustrated, there are a variety of possible predictors that are worth investigating in the future study.

Once a predictor is found, the integration of the algorithms would be simple: to use the predictor to switch the algorithm, or to incorporate the predictor as a feature in the machine learning. For the proof of the latter approach, we extracted the 16 cases where ICD performed the best, and cross-validated the cases by training the CRR model. In 12 diseases, out of 16 diseases, CRR achieved a higher score than ICD, and the Wilcoxon signed rank test confirmed the significance ($p = 0.01$). The result suggests that an appropriate predictor can improve the overall performance, even with a simple machine learning approach.

Related Work

Researchers have been compiling the knowledge source of diseases in a machine-readable format, particularly in the biomedical informatics field: Gene Ontology (Gene Ontology Consortium 2004), the Human Phenotype Ontology

(Robinson and Mundlos 2010), Disease Ontology (Schriml et al. 2012), and Online Mendelian Inheritance in Man (Hamosh et al. 2005) and so on. They include knowledge bases and ontologies that express features of disease in graph structure, and thus, the similarity between concepts can be measured by the similarity in the conceptual relations (Pesquita et al. 2009; Suthram et al. 2010; Mathur and Dinakarpandian 2012; Cheng et al. 2014).

Such a semantic similarity can be applied to the calculation of disease similarity. Mathur and Dinakarpandian (2012) calculated the similarity using genetic relations. Cheng et al. (2014) combined function-based and semantic-based similarities to calculate disease similarity, using the data set in (Pakhomov et al. 2010) as benchmark. Because relationships between diseases and symptoms are often modeled with Bayesian networks (Shwe et al. 1991), they can also be used to calculate disease similarity, in terms of probability.

The limitation of these approaches is their completeness. Although there exist thousands of diseases, most of the researches in the semantic similarity between diseases attempted to test on limited set of diseases against a few targets. Our research selected 80 cases in 1550 diseases, and calculated the similarity against 1550 diseases in the simplified knowledge base, which highly surpasses the settings of the previous attempts. Because there are a variety of diseases, the coverage of the data has decisive importance for the generality of results.

Our further contribution includes a similarity calculation of reasonable quality, utilizing a simplified knowledge base with subjective probability. The performance is achieved without detailed knowledge of diseases, which reduces the cost to develop such a recommendation algorithm for clinical decision support systems.

Conclusion

In this paper, we investigated methods to calculate the similarity between diseases, defining a metric for evaluation of the algorithms. Experimental results suggest that disease similarity is calculated at reasonable quality, even with a superficial calculation on a simplified knowledge base. More precisely, the comparative study suggested that a machine learning approach outperforms the disease classification based approach and the probabilistic approach. The machine learning approach is advantageous, because it can naturally incorporate both the symptomatic knowledge and the etiological knowledge, as well as undocumented preference of user phycisians.

As the future work, it is highly valuable to investigate a factor that predicts a dominant algorithm to measure the similarity for distinct diseases. Additionally, it would also be beneficial to extend the data set to include more diseases, toward further generality of the study.

References

Berner, E. S. 2007. *Clinical Decision Support Systems: Theory and Practice*. Springer Science & Business Media.

Burges, C.; Shaked, T.; Renshaw, E.; Lazier, A.; Deeds, M.; Hamilton, N.; and Hullender, G. 2005. Learning to Rank Using Gradient Descent. In *Proceeding of ICML2005*, 89–96.

Cheng, L.; Li, J.; Ju, P.; Peng, J.; and Wang, Y. 2014. SemFunSim: A New Method for Measuring Disease Similarity by Integrating Semantic and Gene Functional Association. *PLoS ONE* 9(6):e99415.

Gene Ontology Consortium. 2004. The Gene Ontology (GO) Database and Informatics Resource. *Nucleic Acids Research* 32:258–261.

Hamosh, A.; Scott, A. F.; Amberger, J. S.; Bocchini, C. A.; and McKusick, V. A. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research* 33:514–517.

Jaccard, P. 1912. The Distribution of the Flora in the Alpine Zone. *New Phytologist* 11(2):37–50.

Järvelin, K., and Kekäläinen, J. 2000. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *Proceedings of SIGIR '00*, 41–48.

Li, H. 2011. *Learning to Rank for Information Retrieval and Natural Language Processing*. Morgan & Claypool Publishers.

Mathur, S., and Dinakarpandian, D. 2012. Finding disease similarity based on implicit semantic similarity. *Journal of Biomedical Informatics* 45(2):363–371.

Okumura, T.; Tanaka, H.; Omura, M.; Ito, M.; Nakagawa, S.; and Tateisi, Y. 2014. Cost decisions in the development of disease knowledge base : A case study. In *Proceeding of IEEE BIBM 2014*.

Pakhomov, S.; McInnes, B.; Adam, T.; Liu, Y.; Pedersen, T.; and Melton, G. B. 2010. Semantic similarity and relatedness between clinical terms: An experimental study. *AMIA Annu Symp Proc* 2010:572–576.

Pesquita, C.; Faria, D.; Falcao, A. O.; Lord, P.; and Couto, F. M. 2009. Semantic Similarity in Biomedical Ontologies. *PLoS Computational Biology* 5(7):e1000443.

Robinson, P. N., and Mundlos, S. 2010. The human phenotype ontology. *Clinical Genetics* 77(6):525–534.

Schriml, L. M.; Arze, C.; Nadendla, S.; Chang, Y.-W. W.; Mazaitis, M.; Felix, V.; Feng, G.; and Kibbe, W. A. 2012. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research* 40:940–946.

Sculley, D. 2010. Combined regression and ranking. In *Proceeding of KDD-2010*, 979–988.

Shwe, M. A.; Middleton, B.; Heckerman, D.; Henrion, M.; Horvitz, E.; Lehmann, H.; and Cooper, G. 1991. Probabilistic diagnosis using a reformulation of the internist-1/qmr knowledge base. *Methods of information in Medicine* 30(4):241–255.

Suthram, S.; Dudley, J. T.; Chiang, A. P.; Chen, R.; Hastie, T. J.; and Butte, A. J. 2010. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS computational biology* 6(2):e1000662.

World Health Organization. 2011. *ICD-10: International statistical classification of diseases and related health problems.* World Health Organization.

Wu, Z., and Palmer, M. 1994. Verbs Semantics and Lexical Selection. In *Proceedings of ACL '94*, 133–138.