Literature-Based Discovery by an Enhanced Information Retrieval Model

Kazuhiro Seki[1] and Javed Mostafa[2]

[1] Kobe University, Hyogo 657-8501, Japan
seki@cs.kobe-u.ac.jp
[2] Indiana University, Bloomington, IN 47405, USA

Abstract. The massive, ever-growing literature in life science makes it increasingly difficult for individuals to grasp all the information relevant to their interests. Since even experts' knowledge is likely to be incomplete, important findings or associations among key concepts may remain unnoticed in the flood of information. This paper brings and extends a formal model from information retrieval in order to discover those implicit, hidden knowledge. Focusing on the biomedical domain, specifically, gene-disease associations, this paper demonstrates that our proposed model can identify not-yet-reported genetic associations and that the model can be enhanced by existing domain ontology.

Keywords: Hypothesis discovery, Text data mining, Inference network, Implicit association, Gene Ontology.

1 Introduction

With the advance of computer technologies, the amount of scientific knowledge is rapidly growing beyond the pace we could digest. For example, Medline[1]—the most comprehensive bibliographic database in life science—currently indexes over 17 million articles and the number keeps increasing by 1,500–3,000 per day. Given the substantial volume of the publications, it is virtually impossible to deal with the information without the aid of intelligent information processing techniques, such as information retrieval (IR), information extraction (IE), and text data mining (TDM).

In contrast to IR and IE, which find information explicitly stated in documents, TDM aims to discover heretofore unknown knowledge through an automatic analysis on textual data [1]. A pioneering work in TDM, also known as literature-based discovery, was conducted by Swanson in the 1980's. He argued that there were two premises logically connected but the connection had been unnoticed due to overwhelming publications and/or over-specialization. To demonstrate the validity of the basic idea, he manually analyzed numbers of articles and identified logical connections implying a hypothesis that fish oil was effective for clinical treatment of Raynaud's disease [2]. The hypothesis was later supported by experimental evidence.

[1] http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed

V. Corruble, M. Takeda, and E. Suzuki (Eds.): DS 2007, LNAI 4755, pp. 185–196, 2007.

This study is motivated by Swanson's work and attempts to advance the research in literature-based discovery. Specifically, we target implicit associations between genes and hereditary diseases as a test bed. Gene-disease associations are the links between genetic variants and diseases to which the genetic variants influence the susceptibility. For example, BRCA1 is a human gene encoding a protein that suppresses tumor formation. A mutation of this gene increases a risk of breast cancer. Identification of these genetic associations has tremendous importance for prevention, prediction, and treatment of diseases. To this end, we develop a discovery framework by extending the models and techniques developed for IR. Furthermore, we propose the use of domain ontologies for more robust predictions. To demonstrate the effectiveness of the proposed framework, we conduct various evaluative experiments on realistic benchmark data.

2 Related Work

Over two decades, Swanson has argued the potential use of a literature to discover new knowledge that has *implicitly* existed for years but has not been noticed by anybody. His discovery framework is based on a syllogism; i.e., two premises, "A causes B" and "B causes C," suggest a potential association, "A causes C," where A and C do not have a known, explicit relation. Such an association can be seen as a hypothesis testable for verification to produce new knowledge, such as the above-mentioned association between Raynaud's disease and fish oil. For this particular example, Swanson manually inspected two sets of articles concerning Raynaud's disease and fish oil and identified premises that "Raynaud's disease is characterized by high platelet affregability, high blood viscosity, and vasoconstriction" and that "dietary fish oil reduces blood lipids, platelet affregability, blood viscosity, and vascular reactivity," which together suggest a potential benefit of fish oil for Raynaud's patients.

Based on the groundwork, Swanson himself and other researchers developed computer programs to aid hypothesis discovery. The following briefly introduces some of the representative studies.

Weeber et al. [3] implemented a system, called DAD-system, taking advantage of a natural language processing tool. The key feature of their system is that the Unified Medical Language System (UMLS) Metathesaurus[2] was incorporated for knowledge representation and pruning. While the previous work focused on words or phrases appearing in Medline records for reasoning, DAD-system maps them to a set of concepts defined in the UMLS Metathesaurus using MetaMap [4]. An advantage of using MetaMap is that it can automatically collapse different wordforms (e.g., inflections) and synonyms to a single concept. In addition, using *semantic types* (e.g., "Body location or region") under which each Metathesaurus concept is categorized, irrelevant concepts can be excluded from further exploration if particular semantic types of interest are given. This

[2] UMLS is an NLM's project to develop and distribute multi-purpose, electronic knowledge sources and its associated lexical programs.

filtering step can drastically reduce the number of potential associations, enabling more focused knowledge discovery.

Srinivasan [5] developed another system, called Manjal, for literature-based discovery. A key difference of Manjal from the previous work is that it solely relies on MeSH terms assigned to Medline records, disregarding all textual information, so as to study the utility of MeSH terms for hypothesis discovery. Manjal conducts a Medline search for a given concept and extracts MeSH terms from the retrieved articles. Then, according to a predefined mapping, the MeSH terms are grouped into their corresponding UMLS semantic types. Similar to DAD-system, the subsequent processes can be applied only to the concepts under particular semantic types of interest, so as to narrow down the potential associations. Manjal uses the semantic types also for grouping resultant concepts to help its user browse system output. With Manjal, Srinivasan demonstrated that most of the hypotheses Swanson had proposed were successfully replicated.

Despite the prolonged efforts partly mentioned above, however, the research in literature-based discovery can be seen to be at an early stage of development in terms of the models, approaches, and evaluation methodologies. Most of the previous work was largely heuristic without a formal model and their evaluation was limited only on a small number of Swanson's hypotheses. In contrast, this study adapts a formal IR model to literature-based discovery and conducts quantitative experiments based on real-world data.

3 Our Proposed Approach

Focusing on gene-disease associations, we extend a formal IR model, specifically, the inference network [6] for this related but different problem targeting unknown associations. This section details the proposed model and how to estimate the probabilities involved in the model.

In this study, we assume a disease name and known causative genes, if any, as system input. In addition, a target region in the human genome may be specified to limit search space. Given such input, we attempt to predict an unknown causative gene and produce a ranked list of candidate genes.

3.1 An Inference Network for Gene-Disease Associations

In the original IR model, a user query and documents are represented as nodes in a network and are connected via intermediate nodes representing keywords that compose the query and documents. To adapt the model to represent gene-disease associations, we treat disease as query and genes as documents and use two types of intermediate nodes: gene functions and phenotypes which characterize genes and disease, respectively (Fig. 1). An advantage of using this particular IR model is that it is essentially capable of incorporating multiple layers of intermediate nodes. Other popular IR models, such as the vector space models, are not easily applicable as documents and queries are represented by a single layer of the same vocabularies.

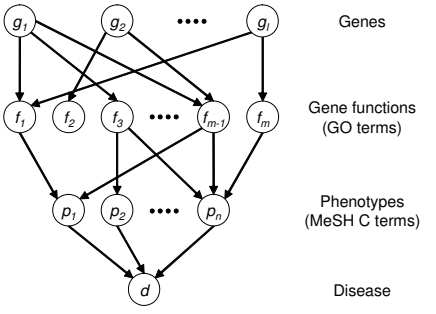

Fig. 1. Inference network for representing gene-disease associations

The network consists of four types of nodes: genes (g), gene functions (f) represented by Gene Ontology (GO) terms,[3] phenotypes (p) represented by MeSH C terms,[4] and disease (d). Each gene node g represents a gene and corresponds to the event that the gene is found in the search for the causative genes underlying d. Each gene function node f represents a function of gene products. There are directed arcs from genes to functions, representing that instantiating a gene increases the belief in its functions. Likewise, each phenotype node p represents a phenotype of d and corresponds to the event that the phenotype is observed. The belief in p is dependent on the belief in f's since phenotypes are (partly) determined by gene functions. Finally, observing certain phenotypes increases the belief in d. As described in the followings, the associations between genes and gene functions ($g \rightarrow f$) are obtained from an existing database, Entrez Gene,[5] whereas both the associations between gene functions and phenotypes ($f \rightarrow p$) and the associations between phenotypes and disease ($p \rightarrow d$) are derived from the biomedical literature.

Given the inference network model, disease-causing genes can be predicted based on the probability defined below.

$$P(d|G) = \sum_i \sum_j P(d|\boldsymbol{p_i}) \times P(\boldsymbol{p_i}|\boldsymbol{f_j}) \times P(\boldsymbol{f_j}|G) \tag{1}$$

Eq. (1) quantifies how much a given set of genes, $G \subseteq \{g_1, g_2, \cdots, g_l\}$, increases the belief in the development of disease d. In the equation, $\boldsymbol{p_i}$ (or $\boldsymbol{f_j}$) is defined as a vector of random variables with i-th (or j-th) element being positive (1) and all others negative (0). By applying Bayes' theorem and some independence assumptions discussed later, we derive

$$P(d|G) \propto \sum_i \sum_j \left(\frac{P(p_i|d)}{P(\bar{p_i}|d)} \times \frac{P(f_j|p_i)P(\bar{f_j}|\bar{p_i})}{P(\bar{f_j}|p_i)P(f_j|\bar{p_i})} \times F(p_i) \times F(f_j) \times P(f_j|G) \right) \tag{2}$$

[3] `http://www.geneontology.org`
[4] `http://www.nlm.nih.gov/mesh`
[5] `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=gene`

where

$$F(p_i) = \prod_{h=1}^{m} \frac{P(\bar{f}_h|p_i)}{P(\bar{f}_h|\bar{p}_i)},\ F(f_j) = \prod_{k=1}^{n} \frac{P(\bar{f}_j)P(f_j|\bar{p}_k)}{P(f_j)P(\bar{f}_j|\bar{p}_k)} \tag{3}$$

The first factor of the right-hand side of Eq. (2) represents the interaction between disease d and phenotype p_i, and the second factor represents the interaction between p_i and gene function f_j, which is equivalent to the odds ratio of $P(f_j|p_i)$ and $P(f_j|\bar{p}_i)$. The third and fourth factors are functions of p_i and f_j, respectively, representing their main effects. The last factor takes either 0 or 1, indicating whether f_j is a function of any gene in G under consideration.

The inference network described above assumes independence among phenotypes, among gene functions, and among genes. We assert that, however, the effects of such associations are minimal in the proposed model. Although there may be strong associations among phenotypes (e.g., phenotype p_x is often observed with phenotype p_y), the model does not intend to capture those associations. That is, phenotypes are attributes of the disease in question and we only need to know those that are frequently observed with disease d so as to characterize d. The same applies to gene functions; they are only attributes of the genes to be examined and are simply used as features to represent the genes under consideration.

3.2 Probability Estimation

Conditional Probability $P(p|d)$. This probability can be seen as a degree of belief that phenotype p is observed when disease d has developed. To estimate the probability, we take advantage of the literature data. Briefly, given a disease name d, a Medline search is conducted to retrieve articles relevant to d and, within the retrieved articles, we identify phenotypes (MeSH C terms) strongly associated with the disease based on chi-square statistics. Given disease d and phenotype p, the chi-square statistic is computed as

$$\chi^2(d,p) = \frac{N(n_{11}\cdot n_{22} - n_{21}\cdot n_{12})^2}{(n_{11}+n_{21})\,(n_{12}+n_{22})\,(n_{11}+n_{12})\,(n_{21}+n_{22})} \tag{4}$$

where N is the total number of articles in Medline, n_{11} is the number of articles assigned p and included in the retrieved set (denoted as R), n_{22} is the number of articles not assigned p and not included in R, n_{21} is the number of articles not assigned p and included in R, and n_{12} is the number of articles assigned p and not in R. The resulting chi-square statistics are normalized by the maximum to treat them as probabilities $P(p|d)$.

Incidentally, for the reason described later, the Medline search is limited to the articles published up to 6/30/2003.

Conditional Probability $P(f|p)$. This probability indicates the degree of belief that gene function f underlies phenotype p. For probability estimation, we adopt the framework similar to the one proposed by Perez-Iratxeta et al. [7].

Unlike them, however, this study focuses on the use of textual data and domain ontologies and investigate their effects for literature-based discovery.

We estimate these probabilities by exploiting the Medline records that were used as the source of the Genetic Association Database (GAD) [8] entries. GAD is a manually-curated archive of human genetic studies, containing pairs of gene and disease that are known to have causative relations. Since each of those Medline records reports experimental evidence indicating causative genetic associations, it is likely to contain descriptions regarding causative associations between phenotypes and gene functions. We can obtain a set of phenotypes (MeSH C terms) associated with a given disease by the same procedure for estimating $P(p|d)$ and can obtain a set of gene functions associated with the gene paired with the disease by consulting the Entrez Gene database. Given the sets of phenotypes and gene functions, our task is to identify which phenotypes and which gene functions have true causative associations and to what degree.

We estimate these associations using three different schemes, i.e., *SchemeK*, *SchemeT*, and *SchemeK+T*. *SchemeK* simply assumes a link between every pair of the phenotypes and gene functions with equal strength, whereas *SchemeT* seeks for evidence in the textual portion of the Medline record to better estimate the strength of associations. Lastly, *SchemeK+T* combines the two schemes by linearly interpolating association scores, $S(f, p)$, described shortly.

SchemeT essentially searches for co-occurrences of gene functions (GO terms) and phenotypes (MeSH terms) in a sliding window, assuming that associated concepts tend to co-occur more often in proximity than unassociated ones. However, a problem is that gene functions and phenotypes are descriptive by nature and may not be expressed in concise GO or MeSH term. To deal with it, we apply the idea of query expansion, a technique used in IR to enrich a query by adding related terms. If GO and MeSH terms are expanded, there is more chance that they could co-occur in text. For this purpose, we use the definitions (or scope notes) of GO and MeSH terms and identify representative terms by inverse document frequencies (IDF), which have been used in IR to quantify term specificity in a document collection. We treat term definitions as documents and define IDF for term t as $\log(N/Freq(t))$, where N denotes the total number of MeSH C (or GO) terms and $Freq(\cdot)$ denotes the number of MeSH C (or GO) terms whose definitions contain term t. Only the terms with high IDF are used as *proxy terms* to represent the original concept, i.e., gene function or phenotype.

Each co-occurrence of the two sets of proxy terms (one representing a gene function and the other representing a phenotype) can be seen as evidence that supports the association between the gene function and phenotype, increasing the strength of their association. We define the increased strength by the product of the term weights, w, for the two co-occurring proxy terms. Then, the strength of the association between gene function f and phenotype p within article a, denoted as $S(f, p, a)$, can be defined as the sum of the increases for all co-occurrences of the proxy terms in a. That is,

$$S(f,p,a) = \sum_{(t_f,t_p,a)} \frac{w(t_f) \cdot w(t_p)}{|Proxy(f)| \cdot |Proxy(p)|} \tag{5}$$

where t_f and t_p denote any terms in the proxy term sets for f and p, respectively, and (t_f, t_p, a) denotes a set of all co-occurrences of t_f and t_p within a. The product of the term weights is normalized by the proxy size, $|Proxy(\cdot)|$, to eliminate the effect of different proxy size. (Note that a larger proxy size generally produces a greater numerator.) As term weight w, we used the TF·IDF weighting scheme. For term t_p for instance, we define $\mathrm{TF}(t_p)$ as $1 + \log Freq(t_p, Def(p))$, where $Def(p)$ denote p's definition and $Freq(t_p, Def(p))$ denotes the number of occurrences of t_p in $Def(p)$.

The association scores, $S(f, p, a)$, are computed for each GAD entry by either *SchemeK* or *SchemeT* and are accumulated over all entries to estimate the associations between f's and p's, denoted as $S(f, p)$. Based on the associations, we define probability $P(f|p)$ as $S(f, p)/\sum_p S(f, p)$.

A possible shortcoming of the approach described above is that the obtained associations $S(f, p)$ are symmetric despite the fact that the network in Fig. 1 is directional. However, since it is well-known that an organism's genotype (in part) determines its phenotype, we assume the estimated associations between gene functions and phenotypes to be directed from the former to the latter.

Enhancing $P(f|p)$ by Domain Ontology. The proposed framework may not be able to establish true associations between gene functions and phenotypes for various reasons, e.g., the amount of training data may be insufficient. Those true associations may be uncovered using the structure of MeSH and/or GO. MeSH and GO have a hierarchical structure[6] and those located nearby in the hierarchy are semantically close to each other. Taking advantage of these properties, we enhance the estimated probabilities $P(f|p)$ as follows.

Let A denote the matrix whose element a_{ij} is probability estimate $P(f_j|p_i)$ and A' denote the enhanced or updated matrix. Then, A' is formalized as $A' = W_p A W_f$, where W_p denotes an $n \times n$ matrix with element $w_p(i, j)$ indicating a proportion of a probability to be transmitted from phenotypes p_j to p_i. Similarly, W_f is an $m \times m$ matrix with $w_f(i, j)$ indicating a proportion transmitted from gene functions f_i to f_j. This study experimentally uses only direct child-to-parent and parent-to-child relations and defines the weight function $w_p(i, j)$ as

$$w_p(i,j) = \begin{cases} 1 & \text{if } i = j \\ \dfrac{1}{\#\text{ of children of } p_j} & \text{if } p_i \text{ is a child of } p_j \\ \dfrac{1}{\#\text{ of parents of } p_j} & \text{if } p_i \text{ is a parent of } p_j \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Eq. (6) means that the amount of probability is equally split among its children (or parents). Similarly, $w_f(i, j)$ is defined by replacing i and j in the right-hand side of Eq. (6). Note that this enhancement can be iteratively applied to take advantage of more distant relationships than children/parents.

[6] To be precise, GO's structure is directed acyclic graph, allowing multiple parents.

Table 1. Number of gene-disease associations in the benchmark data

	Cancer	Cardio-vascular	Immune	Metabolic	Psych	Unknown	Total
Training	258	305	376	313	172	864	2,288
Test	45	36	61	23	12	80	257
Total	303	341	437	336	184	944	2,545

4 Empirical Evaluation

4.1 Benchmark Data

To evaluate the validity of the proposed approach, we implemented a prototype system and conducted various experiments on the benchmark data sets created from GAD. The following details the creation of the benchmark data.

1. Associate a gene-disease pair from each GAD entry with the publication date of the article from which the entry was created. The date can be seen as the time when the causative relation became public knowledge.
2. Group gene-disease pairs based on disease names. (As GAD deals with complex diseases, a single disease may be paired with multiple genes.)
3. For each pair of a disease and its causative genes,
   (a) Identify the gene whose relation to the disease was most recently reported based on the publication date. If the date is on or after 7/1/2003, the gene will be used as the target (i.e., new knowledge), and the disease and the rest of the causative genes will be used as system input (i.e., old knowledge). In other words, the target-input pair will be an instance composing test data. If the date is before 7/1/2003, the pair of the disease and the gene is added to training data.
   (b) Remove the most recently reported gene from the set of causative genes and return to step (3a).

The separation by publication dates ensures that a training phase does not use new knowledge in order to simulate gene-disease association discovery. The particular date was arbitrarily chosen by considering the size of the resulting data sets. Table 1 shows the number of gene-disease associations in the resulting data sets under six disease classes defined in GAD. In the following experiments, the cancer class is used for system development and parameter tuning.

4.2 Experimental Setup

Given input (disease name d, known causative genes C, and target region r), the system computes the probability $P(d|G)$ as in Eq. (3) for each candidate gene g located in r, where G is C plus g. For instance, d, C, and r might be `hepatocellular carcinoma`, {`APC`,`IL1`}, and `8q24`, respectively. The candidate genes are then output in a decreasing order of their probabilities.

Table 2. System performance in AUC for each disease class. The figures in the parentheses indicate percent increase/decrease relative to *SchemeK*.

	Cardio-vascular	Immune	Metabolic	Psych	Unknown	Overall
K	0.707	0.612	0.681	0.628	0.684	0.661
T	0.731 (3.4%)	0.611 (-0.2%)	0.614 (-9.9%)	0.667 (6.2%)	0.761 (11.2%)	0.686 (3.8%)
K+T	0.697 (-1.4%)	0.656 (7.2%)	0.682 (0.1%)	0.702 (11.8%)	0.743 (8.5%)	0.699 (5.9%)

As evaluation metrics, we use *area under the ROC curve* (AUC) for its attractive property as compared to the F-score measure (see Fawcett [9] for details). ROC curves are two dimensional measure for system performance with x axis being true positive proportion (TPP) and y axis being false positive proportion (FPP). TPP is defined as TP/(TP+FN), and FPP as FP/(FP+TN), where TP, FP, FN, and FP denote the number of true positives, false positives, false negatives, and false positives, respectively. AUC takes a value between 0 and 1 with 1 being the best. Intuitively AUC indicates the probability that a gene randomly picked from positive set is scored more highly than one from negative set.

Probabilities $P(f|p)$ were pre-computed using the training data. Then the test data in the cancer class were used to determine several parameters for each scheme, including the number of Medline articles as the source of phenotypes (n_m), threshold for chi-square statistics (t_c), threshold for IDF to determine proxy terms (t_t), and window size for co-occurrences (w_s). For example, for *SchemeT*, they were set as n_m=400, t_c=2.706, t_t=4.0, and w_s=100 (words) by testing numbers of their combinations.

4.3 Results and Discussions

Overall Performance. With the optimal parameter settings identified with the cancer class, the system was applied to all the other disease classes. Table 2 summarizes the system performance in AUC.

All the schemes achieved significantly higher AUC than 0.5 (which corresponds to a random guess), indicating the validity of the general framework using the inference network for discovering implicit associations. For individual disease classes, it is observed that *SchemeT* yielded the best AUC for the *Cardiovascular* and *Unknown* classes and *SchemeK+T* for the others. Overall, *SchemeK+T* works the best, followed by *SchemeT*. The difference between *SchemeK+T* and *SchemeK* is significant ($p < 0.01$), which proves the benefit of textual information.

Enhancing $P(f|p)$ by Domain Ontology. Section 3.2 discussed that ontology could be exploited to enhance probability estimates $P(f|p)$. In brief, considering parent-to-child (P-to-C) and child-to-parent (C-to-P) relations between two concepts defined in MeSH and GO, one could deduce associations between

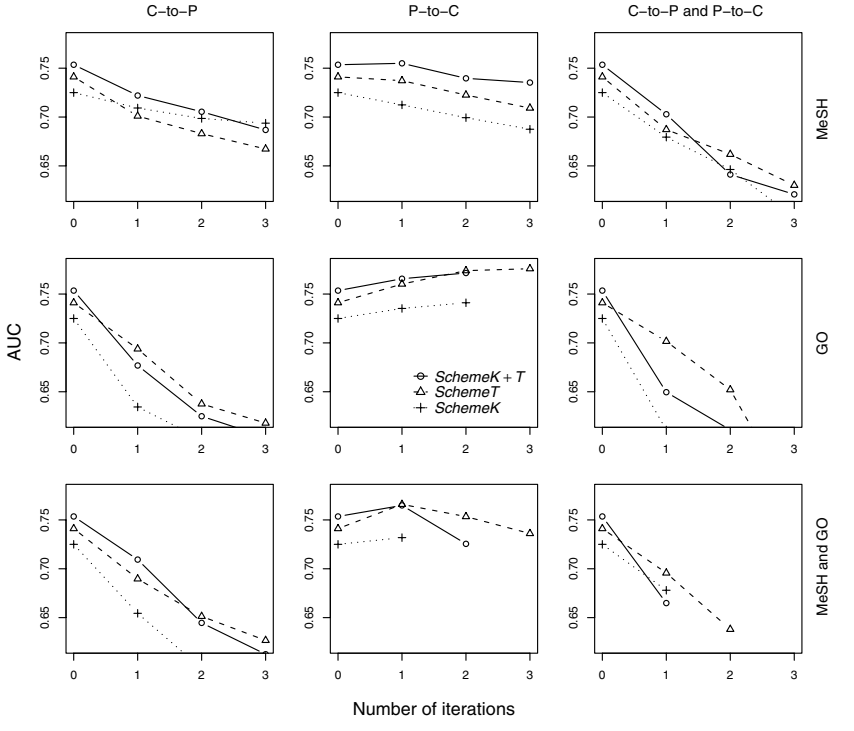

Fig. 2. Transitions of AUC for different source and direction of relations

the concepts that were *not* observed in the training phase. Again, using the cancer class data, we investigated an effective use of domain ontologies.

Eq. (6) was defined to use both MeSH and GO and both P-to-C and C-to-P relations. However, it is expected that each knowledge source and each direction of the relations would have different effects on the outcome. To determine the best strategy, we compare the combination of the following alternative settings: only MeSH, only GO, or both for the source of the semantic relations, and only P-to-C, only C-to-P, or both for the direction of the relations. Because these two properties are independent, there are $3 \times 3 = 9$ different combinations to be examined. Fig. 2 shows plots for these combinations, where x and y axes represent the number of iterations and AUC, respectively. Note that, due to the limitation of computer memory used for this experiment, we could iterate the computation only once or twice for some cases.

Contrary to our expectation, the use of the ontologies rather deteriorated AUC for many cases. Especially, when C-to-P relations were considered (the left and right columns), AUC dropped as the number of iterations increased regardless of the scheme used. On the other hand, when GO and only P-to-C relations were used (the center and bottom middle), AUC mildly improved at

Table 3. System performance in AUC after enhancing probability estimates with only parent-to-child relations in GO hierarchy. The figures in the parentheses indicate percent increase/decrease relative to the corresponding cells in Table 2.

	# of iterations	Cardio-vascular	Immune	Metabolic	Psych	Unknown	Overall
K	1	0.707 (0.0%)	0.601 (-1.8%)	0.702 (3.0%)	0.672 (7.0%)	0.706 (3.2%)	0.673 (1.8%)
T	1	0.727 (-0.5%)	0.609 (-0.4%)	0.618 (0.6%)	0.750 (12.4%)	0.774 (1.7%)	0.695 (1.4%)
	2	0.726 (-0.6%)	0.613 (0.4%)	0.633 (3.1%)	0.768 (15.0%)	0.782 (2.7%)	0.703 (2.5%)
K+T	1	0.703 (0.9%)	0.631 (-3.8%)	0.698 (2.3%)	0.765 (8.9%)	0.763 (2.7%)	0.708 (1.3%)

least at the first iteration. These results suggest that the associations between gene functions and phenotypes could be safely enhanced only downwards in the hierarchies. Among the two plots, using only GO hierarchy (the center) shows constant improvement of AUC with the number of iterations, whereas the other (the bottom middle) gradually declines from the second iteration. The best AUC (=0.776) was achieved with *SchemeT* after three iterations using only GO and P-to-C relations.

Based on these observations, the same strategy (i.e., GO with P-to-C) was applied to all the other disease classes; The results are summarized in Table 3.

As shown, the system performance more or less improved except for the *Cardiovascular* and *Immune* classes. Overall, AUC marginally increased irrespective of the schemes. After applying two times of iterations to *SchemeT*, it further improved to 0.703. (It could not be applied to the other two due to the memory limitation.) These experiments verify that the strategy of using P-to-C relations in the GO hierarchy is generally effective in other types of diseases and that system performance slightly but steadily increases with the number of iterations. The improvement of *SchemeT* is statistically significant at the 5% level.

In the experiments above, considering the MeSH hierarchy was found harmful in enhancing $P(f|p)$. It may have been caused by a possible difference in the nature of the MeSH and GO hierarchies. To investigate, we compared their organizational structures (e.g., the number of children per node) but were not able to find notable difference in this regard. Another cause of the problem may be possible spurious phenotypes associated with a query disease. Remember that while GO terms are obtained from Entrez Gene given a candidate gene (i.e., a simple database lookup), MeSH terms are harvested from Medline search results with a disease name being a query, assuming that MeSH terms annotated with the retrieved articles are representative phenotypes of the disease. Thus, some of those MeSH terms may not be associated with the disease at all. Enhancing associations based on those spurious phenotypes, if any, would degrade system prediction. More work needs to be done to determine the benefit of MeSH.

5 Concluding Remarks

This paper explored a novel discovery framework targeting implicit gene-disease associations and proposed an extension of IR models/techniques in conjunction with domain-specific resources, such as the literature, gene database, and ontology. To examine the validity of the framework, we created realistic benchmark data, where old and new knowledge were carefully separated to simulate knowledge discovery. The key findings identified by empirical observations include that a) the consideration of textual information improved system prediction by 5.9% in AUC over simply relying on co-annotations of keywords, and b) semantic relations defined in domain ontologies could be leveraged to enhance probability estimates, where MeSH were found rather harmful in the current scheme.

For future work, we plan to investigate more sophisticated schemes, e.g., the semantic distance [10], in propagating the probabilities $P(f|p)$. In addition, we would like to compare the proposed framework with the previous work and with other IR models so as to study the characteristics/advantages of our model.

Acknowledgments. This work was partially supported by the Artificial Intelligence Research Promotion Foundation grant #18AI-255, the Nakajima Foundation, the Japanese Ministry of Education, Culture, Sports, Science and Technology, and the NSF grant #0549313.

References

1. Hearst, M.A.: Untangling text data mining. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 3–10 (1999)
2. Swanson, D.R.: Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine 30(1), 7–18 (1986)
3. Weeber, M., Klein, H., de Jong-van Berg den, L.T.W., Vos, R.: Using concepts in literature-based discovery: simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. Journal of the American Society for Information Science and Technology 52(7), 548–557 (2001)
4. Aronson, A.R.: Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In: Proceedings of American Medical Informatics 2001 Annual Symposium, pp. 17–21 (2001)
5. Srinivasan, P.: Text mining: generating hypotheses from Medline. Journal of the American Society for Information Science and Technology 55(5), 396–413 (2004)
6. Turtle, H., Croft, W.B.: Evaluation of an inference network-based retrieval model. ACM Transactions on Information Systems 9(3), 187–222 (1991)
7. Perez-Iratxeta, C., Wjst, M., Bork, P., Andrade, M.: G2D: a tool for mining genes associated with disease. BMC Genetics 6(1), 45 (2005)
8. Becker, K.G., Barnes, K.C., Bright, T.J., Wang, S.A.: The genetic association database. Nature Genetics 36, 431–432 (2004)
9. Fawcett, T.: ROC graphs: Notes and practical considerations for researchers. Technical Report HPL-2003-4, HP Laboratories (2004)
10. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research 11, 95–130 (1999)