Hypothesis Generation and Ranking Based on Event Similarities

Taiki Miyanishi
Graduate School of
Engineering, Kobe University
miyanishi@ai.cs.kobe-u.ac.jp

Kazuhiro Seki
Organization of Advanced
Science and Technology
Kobe University
seki@cs.kobe-u.ac.jp

Kuniaki Uehara
Graduate School of
Engineering, Kobe University
uehara@kobe-u.ac.jp

ABSTRACT

Accelerated by the technological advances in the domain, the size of the biomedical literature has been growing rapidly. As a result, it is not feasible for individual researchers to comprehend and synthesize all the information related to their interests. Therefore, it is conceivable to discover hidden knowledge, or *hypotheses*, by linking fragments of information independently described in the literature. In fact, such hypotheses have been reported in the literature mining community; some of which have even been corroborated by experiments. This paper mainly focuses on hypothesis ranking and investigates an approach to identifying reasonable ones based on semantic similarities between events which lead to respective hypotheses. Our assumption is that hypotheses generated from semantically similar events are more reasonable. The validity of our approach is demonstrated in comparison with those based on term frequencies, often adopted in the related work.

Categories and Subject Descriptors

I.7 [**Document and Text Processing**]; J.3 [**Life and Medical Sciences**]

General Terms

Algorithms, Measurement, Design

Keywords

Biomedical Text Mining, Hypotheses Ranking, Ontology

1. INTRODUCTION

The biomedical literature has been rapidly growing at a rate of several thousand papers per week, which makes it infeasible for individual researchers to comprehend all the information related to their interests [1, 3, 9]. For this reason, it is conceivable that much potential knowledge, or *hypotheses*, is remaining undiscovered in the large amount of data. In fact, such hypotheses have been reported in the literature mining community; some of which have even been corroborated by experiments [4, 5, 14, 20, 21, 22].

As the pioneering work for biomedical hypotheses discovery, Swanson [21] predicted that fish oil would be effective for the treatment of Raynaud's disease by manually investigating and linking multiple information independently described in the literature. This hypothesis was later validated by Digiacomo [4]. Following Swanson, other research groups reexamined and extended his work on hypothesis discovery in an attempt to automatically identify promising hypotheses [6, 24, 19, 16, 25]. However, their methods typically require manual intervention to generate hypotheses and do not have a mechanism to properly deal with low frequency terms/concepts since they are basically based on term frequencies.

In hypothesis discovery, it is indeed possible that there is promising, hidden knowledge derived from infrequent terms and, due to their infrequencies, those hypotheses can be easily overlooked irrespective of their importance. Thus, a discovery framework should not be dependent (solely) on term frequencies. Also, if we do not use term frequencies, it is crucial to focus on only significant topics closely associated with the main theme of an article since considering many infrequent terms indiscriminately would lead to numerous, meaningless hypotheses.

Motivated by the background, the aim of this paper is to investigate an automatic hypothesis generation framework with a focus on a ranking function which considers not term frequencies but semantic similarity between two events that lead to a hypothesis. Our main assumption is that semantically similar events yield a more reasonable hypothesis.

The rest of this paper is organized as follows. Section 2 summarizes the previous work most related to this paper. Section 3 details our developed framework for hypothesis generation and ranking based on event similarities. Section 4 evaluates our hypothesis ranking functions in comparison with term frequency-based functions. Finally, Section 5 concludes with brief summary and possible future directions.

2. RELATED WORK

Swanson's framework for hypothesis discovery is based on the idea, so called the *ABC* syllogism. It discovers an implicit connection between two concepts, such as "A causes C," when it is well acknowledged that "A causes B" and "B causes C" while A and C do not have a relationship explicitly reported in the literature.

It may sound trivial to predict the potential relationship

SAC'10 March 22-26, 2010, Sierre, Switzerland.

between A and C provided that the relationships between A and B and between B and C exist. However, it is not necessarily the case for humans if the two relations are found in two different literatures representing two different specialties or if they reside in one literature which, however, is too large to look through for individuals. Both situations are conceivable in the biomedical domain given the overwhelming publications and various specialties.

Motivated by Swanson's work, several other researchers, as well as Swanson himself [23], have developed computer systems to aid hypothesis discovery [6, 8, 12, 19, 25, 26, 27]. Here we focus on two important attempts most related to our work described in this paper.

Weeber [25] implemented a system, called DAD-system, to support hypothesis discovery by taking advantage of a natural language processing (NLP) tool. The key difference of his system from the others was that Weeber used the Unified Medical Language System (UMLS) Metathesaurus[1] for representing and filtering concepts. Unlike Swanson [23], each sentence in textual portions of MEDLINE records was mapped to the concepts defined in the UMLS Metathesaurus by the MetaMap program [2] instead of extracting words or phrases from the sentence. For example, MetaMap converts an input sentence "Platelet aggregation is known to be high in patients with Raynaud's syndrome." to five concepts: "Platelet aggregation," "Known," "High," "Patients," and "Raynaud's disease," where word variants (e.g., singular vs. plural, synonyms, inflection) can be mapped to a single concept. The identified concepts are then filtered based on their semantic categories; each concept in Metathesaurus is assigned one or more semantic categories called *semantic types*, such as "Body location or region," "Vitamin," and "Physiologic function." Given a set of semantic types of particular interest, this filtering step could drastically reduce the number of potential concepts to a manageable size.

Srinivasan [19] developed another system, called Manjal,[2] for hypothesis discovery. In contrast to the previous work which mainly used the textual portion of MEDLINE records (i.e., titles and abstracts), she focused solely on Medical Subject Headings (MeSH) terms assigned to MEDLINE records in conjunction with the UMLS semantic types and investigated their effects for discovering implicit associations. MeSH is a thesaurus maintained by National Library of Medicine (NLM) for indexing articles in life sciences. Given a starting concept A, the Manjal system conducts a MEDLINE search for A and extracts MeSH terms from the retrieved MEDLINE records as B concepts. Then, the B concepts are grouped into the corresponding UMLS semantic types according to a predefined mapping. Similar to the approach by Weeber [25], the subsequent processes can be limited only to the concepts under the specific semantic types of interest, so as to narrow down the potential pathways. A difference from Weeber's approach is that, besides it used only MeSH terms, she used the TFIDF term weighting scheme [18] to rank B or C concepts so as to make potentially important connections more noticeable to the user. TFIDF is an abbreviation of "term frequency-inverse document frequency", originally developed for information retrieval to quantify the importance of a term to specify a document containing the term.

For hypothesis generation, we will take an approach similar to Weeber's but utilizes our semantic-based ranking functions described shortly to give an order to numbers of generated hypotheses. The ranking functions will be evaluated in comparison with frequency-based functions, such as TFIDF.

3. PROPOSED FRAMEWORK

This section first describes how to generate hypotheses based on a biomedical entity network extracted from the literature. Then, a new concept of *hypothesis reasonability* is introduced to identify promising hypotheses. We instantiate two variants of ranking functions based on event similarities along with two frequency-based ranking functions typically used in the previous work.

3.1 Hypothesis Generation

To generate hypotheses, we first construct a biomedical entity network, which requires named entity recognition and relationship extraction. For these processes, we take an approach similar to Weeber [25]. Weeber used MetaMap program [2] to obtain UMLS concepts and extracted relationships between concepts based on their co-occurences in titles and abstracts of MEDLNE records. Different from Weeber, who used all concepts MetaMap output, we use only concepts with the highest mapping scores to raise the precision of entity recognition. Also, we use only titles to avoid producing many meaningless hypotheses. We chose titles because they can be seen as concise, high quality summary of the articles as reported in the functional genomics domain [7]. Another reason to use titles is our simplified assumption to regard co-occurring concepts as related. The relation between two co-occurring concepts may be described either affirmatively or negatively, which ideally needs to be determined through syntactic analysis. However, article titles were found often affirmative by our informal observation and thus more suited to our current event extraction scheme.

Another difference from Weeber's approach is that we consistently apply a set of UMLS semantic types for filtering, whereas Weeber arbitrarily used two different sets of semantic types for intermediate concepts and for end concepts, respectively.

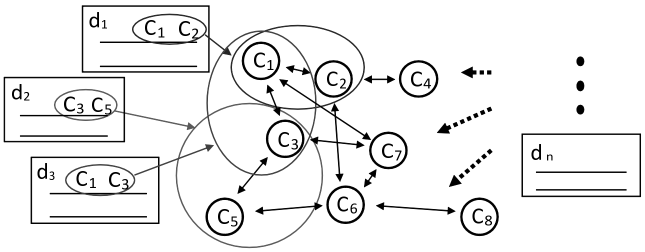

Figure 1: Entity network constructed from the literature.

We regard a co-occurrence of two concepts in an article title as an event disregarding the type of the event for simplicity. Then, by merging common concepts of the extracted events, an entity network can be constructed. Figure 1 shows an example of an entity network consisting of such binary relationships extracted from the biomedical articles, d_1, d_2, d_3, …, d_n. In Figure 1, event c_1-c_2 (composed of the two

[1] UMLS is an NLM's project to develop and distribute multipurpose, electronic knowledge sources and its associated lexical programs. http://www.nlm.nih.gov/research/umls/

[2] http://sulu.info-science.uiowa.edu/Manjal.html

concepts c_1 and c_2) was extracted from d_1, event c_3-c_5 was extracted from d_2, event c_1-c_3 was extracted from d_3, and so on. We can discover a potentially new relationship c_1-c_5 by following the path of the events c_1-c_3 and c_3-c_5. It is only possible to discover these indirect relationships by gathering the information found in separate articles, d_1, d_2, and d_3, together.

More formally, hypothesis generation can be performed by exploring the entity network with a given starting point. The search is terminated when another concept of interest given by a user is found or when exploration reaches a specified depth. The discovered paths between starting and ending points are interpreted as hypotheses except for closed chains in the network. We call the starting, ending, and intermediate points A-term, C-term, and B-term, respectively. Figure 2 shows the resulting pathways aligned by A, B, and C-terms based on Figure 1, where node c_1 was used as a starting point for search.

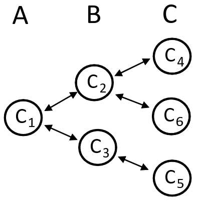

Figure 2: Example of *A*,*B*, and *C*-terms.

3.2 Reasonability of Hypothesis

The generated hypothesis becomes new knowledge after verifying it through actual experiments, which are often costly to conduct. Our purpose is to identify reasonable hypotheses that are more likely to lead to new knowledge among automatically deduced hypotheses. As described, we call a binary relationship an event, and a hypothesis is a new event derived from more than two events sharing a common entity. To measure the reasonability of a hypothesis, we make an assumption that a reasonable hypothesis would be generated from semantically similar events. This assumption is based on the intuition that a hypothesis generated from dissimilar events has a semantic gap that is difficult to interpret. A hypothesis generated from similar events, on the other hand, would be more logical or more easily to understand, resulting in more reasonable hypotheses. Then, the question is how to measure the similarity of events to quantify the reasonability of a hypothesis. For this purpose, we take advantage of MeSH terms.

MeSH terms are assigned to MEDLINE records to characterize each article. Each MEDLINE record is assigned approximately ten MeSH terms by hand. The 2009 version of MeSH contains a total of 25,186 subject headings, also known as descriptors. MeSH descriptors are arranged in both an alphabetic and a hierarchical structure. The terms at the most general level of the hierarchical structure are very broad headings. More specific headings are found at narrower levels of the eleven-level hierarchy [13].

Here, a potential problem regarding MeSH terms may be that they characterize not an event but an article with which they are assigned. As described, however, we extract events only from the main part of an article, namely, the article title, which we expect to be well represented by MeSH terms.

In the following, we first introduce concept similarity based on MeSH terms and then extend it to event similarity.

3.3 Concept Similarity

There is a variety of semantic similarity measures defined on a thesaurus like MeSH [10, 11, 15, 17]. Among them, we adopt the measure proposed by Seco [17] because his measure depends only on the structure of a thesaurus without term/concept frequencies.

Seco assumed that, in a given thesaurus, concepts with many hyponyms convey less information than those with low hyponyms. If concepts are the most specific in the thesaurus (i.e., leaf nodes), the information they provide is maximum. Based on the assumption, the semantic similarity between two concepts, m_1 and m_2, is expressed using Information Content (IC) defined as

$$\mathrm{sim}(m_1, m_2) = \max_{m \in S(m_1, m_2)} \mathrm{IC}(m) \qquad (1)$$

$$\mathrm{IC}(m) = 1 - \frac{\log(\mathrm{hypo}(m) + 1)}{\log(\mathrm{N_s})} \qquad (2)$$

where $\mathrm{IC}(m)$ is the IC value of a MeSH term m, $S(m_1, m_2)$ is a set of concepts that subsumes both m_1 and m_2 in the thesaurus, $\mathrm{hypo}(m)$ is the number of hyponyms of m, and $\mathrm{N_s}$ is the total number of concepts in the thesaurus. The denominator in Equation (2), which is equivalent to the value of the least informative concept, serves as a normalizing factor to ensure that IC values are in [0,1]. This formulation guarantees that IC decreases monotonically with the generality of a concept. Moreover, IC of the imaginary top node of a thesaurus becomes 0.

3.4 Event Similarity

The concept similarity defined in Equation (1) is extended to event similarity, which we regard as the reasonability of a hypothesis. The relationship between the hypothesis reasonability and the event similarity is illustrated in Figure 3. The reasonability of the hypothesis linking c_1-c_5 is the similarity between events c_1-c_3 and c_3-c_5, which is higher than that of the hypothesis linking c_1-c_6 since the similarity between events c_1-c_3 and c_3-c_5 is higher than that between events c_1-c_2 and c_2-c_6. The following describes two simple instantiations of event similarity extended from the concept similarity.

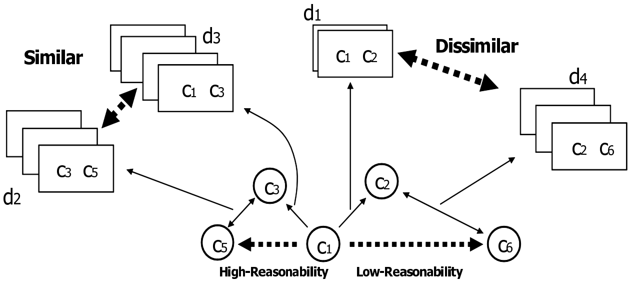

Figure 3: Relation between event similarity and hypothesis reasonability.

3.4.1 Event Similarity by Concept Similarity Averaging

A straightforward extension from the concept similarity would be to take an average of the similarities between all the combinations of the concepts representing two events. This similarity, or the reasonability of a resulting hypothesis, can be defined as

$$\mathrm{R_{avg}}(e_i, e_j) = \frac{1}{|\mathbf{M}_i||\mathbf{M}_j|} \sum_{m_k \in \mathbf{M}_i} \sum_{m_l \in \mathbf{M}_j} \mathrm{sim}(m_k, m_l) \qquad (3)$$

where $\mathbf{M}_i$ and $\mathbf{M}_j$ are sets of MeSH terms corresponding to events e_i and e_j, respectively. A set of MeSH terms $\mathbf{M}$ is formed by aggregating MeSH terms from the MEDLINE records from which an event is extracted. A shortcoming of this similarity $\mathrm{R_{avg}}$ is that it considers the similarities even between dissimilar concepts as we will discuss in Section 4.

3.4.2 Event Similarity by Nearest Concept Similarity Averaging

This definition of event similarity, $\mathrm{R_{max}}$, intends to deal with the problem of $\mathrm{R_{avg}}$ briefly mentioned above by focusing only on the most similar concepts. For every concept representing an event e_i, we identify the most similar concept representing another event e_j and take the average of the similarities. Then, we switch e_i and e_j and compute the average. $\mathrm{R_{max}}$ is defined as the sum of the averages to make it symmetric.

$$\mathrm{R_{max}}(e_i, e_j) = \frac{1}{|\mathbf{M}_i|} \sum_{m_k \in \mathbf{M}_i} \max_{m_l \in \mathbf{M}_j} \mathrm{sim}(m_k, m_l) + \frac{1}{|\mathbf{M}_j|} \sum_{m_l \in \mathbf{M}_j} \max_{m_k \in \mathbf{M}_i} \mathrm{sim}(m_k, m_l) \qquad (4)$$

3.4.3 Event Similarity by TFIDF

To be compared with the above two semantic-based ranking functions, we also introduce a frequency-based event similarity adopting the often-used TFIDF term weighting scheme. An event may be extracted from multiple articles, and thus, it may have some duplicated MeSH terms. We regard the number of duplications of a MeSH term as its term frequency and the number of articles indexed with the MeSH term as its document frequency. Then, each event e_i can be represented as a MeSH term vector weighted by TFIDF. That is,

$$e_i = (w_{i1}, w_{i2}, \cdots, w_{in})$$

where $w_{ij} = v_{ij} / \max_k(v_{ik})$ is the normalized TFIDF for MeSH term m_{ij} and v_{ij} is defined as $n_{ij} \times \log(N/n_j)$. N is the total number of documents, n_j is the number of documents indexed with MeSH term m_j, and n_{ij} is the number of duplications for MeSH term m_j extracted for event e_i.

Using the MeSH vectors, the similarity between events e_i and e_j can be computed by cosine similarity. This definition resembles the one proposed by Srinivasan [19].

3.4.4 Event Similarity by Event Frequencies

We define yet another event similarity measure based on event frequencies as

$$\mathrm{R_{freq}}(e_i, e_j) \quad = \quad \sqrt{freq(e_i) \times freq(e_j)} \qquad (5)$$

where $freq(e)$ denotes the frequency of an event. The intuition behind this is that the hypothesis supported by highly frequent events is reasonable.

4. EVALUATION

4.1 Experimental Settings

We used the Swanson's discovered hypothesis in 1986 that fish oil is effective for the treatment of Raynaud's disease, so as to evaluate our generated hypotheses. In short, Raynaud's disease is characterized by blood viscosity, platelet aggregability, and vascular reactivity, and fish oil is able to ease these symptoms. We examined if our semantic-based reasonability measures can make these associations prominent.

To be precise, we looked at the biomedical literatures published between 1960 and 1985 and used their titles for constructing an entity network. Given fish oil as a starting concept, the hypotheses generated from the network were ranked using the reasonability measures defined in Section 3.4. If hypotheses with correct pathways (i.e., blood viscosity, platelet aggregation, vascular reactivity, or the like) are ranked higher by the semantic similarity-based reasonabilities than by the frequency-based reasonabilities, it suggests that the former would be useful for identifying important hypotheses that may be overlooked otherwise.

Following Weeber [25], nine UMLS semantic types below were used for restriction at the relationship extraction step: Biologic Function, Cell Function, Disease or Syndrome, Lipid, Molecular Function, Organ or Tissue Function, Organism Function, Pathologic Function, and Physiologic Function. Additionally, a UMLS concept, Blood Viscosity [Laboratory or Test Result], was substituted with Blood Viscosity [Physiologic Function] since the UMLS Semantic type "Laboratory or Test Result" is not relevant for this experiment. As a result, we obtained an entity network composed of 15,774 nodes and 193,165 edges.

4.2 Results and Discussion

Given fish oil as the A-term, 13,677 hypotheses were found by searching the entity network at the depth of two. Among them, there were eight hypotheses whose C-term was Raynaud's disease. Table 1 shows these hypotheses represented by A, B, and C-terms, sorted alphabetically by their B-terms. Note that "Primary Raynaud's" and "Paroxysmal digital cyanosis" are synonyms of Raynaud's disease. In Table 1, "Blood Viscosity" is a meaningful concept which legitimately connects fish oil to Raynaud's disease as mentioned above. In addition, since "Atheromatosis" and "Peripheral vascular disease" are related to platelet aggregation and blood viscosity, the hypotheses H_1, H_6, and H_7 are reasonable, too. On the other hand, "Development" and "Suppression" are too general to be useful in order to link fish oil and Raynaud's disease. Thus, the hypotheses H_5 and H_8 are not considered reasonable.

We ranked all the 13,677 hypotheses using the reasonability measures described in Section 3.4. Figures 4 and 5 plot their rankings (shown in proportion to the total number of hypotheses) of the reasonable and unreasonable hypotheses, respectively. For the former, the higher the hypotheses are ranked (i.e., having smaller values), the better the reasonability measures are. For the latter, conversely, good reasonability measures should rank the hypotheses lower (i.e., having larger values). The rightmost points in Figures 4 and 5 are the respective average rankings obtained by different reasonability measures. To remind, $\mathrm{R_{avg}}$ is the reasonability measure defined as the average of the similarities for all

Table 1: Generated hypotheses

ID	Reasonable	*A*-term	*B*-term	*C*-term
H_1	Yes	Fish Oil	Atheromatosis	Raynaud Disease
H_2	Yes	Fish Oil	Blood Viscosity	Paroxysmal digital cyanosis
H_3	Yes	Fish Oil	Blood Viscosity	Primary Raynaud's
H_4	Yes	Fish Oil	Blood Viscosity	Raynaud Disease
H_5	No	Fish Oil	Development	Paroxysmal digital cyanosis
H_6	Yes	Fish Oil	Peripheral vascular disease	Paroxysmal digital cyanosis
H_7	Yes	Fish Oil	Peripheral vascular disease	Raynaud Disease
H_8	No	Fish Oil	Suppression	Paroxysmal digital cyanosis

combinations of the concepts representing two events, and R_{max} is the average of the most similar concepts. R_{tfidf} is the cosine similarity of the concepts weighted by TFIDF. R_{freq} is the geometric mean of the two event frequencies.

We first discuss the overall rankings of reasonable and unreasonable hypotheses. As mentioned, the rankings of the reasonable hypotheses in Figures 4 are considered better when they have lower values. On average, R_{max} performed the best followed by R_{tfidf}, R_{avg}, and R_{freq}. For the average rankings of the unreasonable hypotheses, Figures 5 shows that R_{avg} and R_{max} were able to rank them lower (i.e., higher values) than R_{tfidf} and R_{freq}.

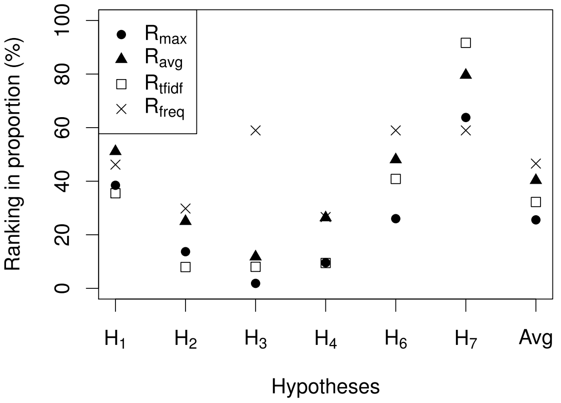

Figure 4: Rankings of reasonable hypotheses.

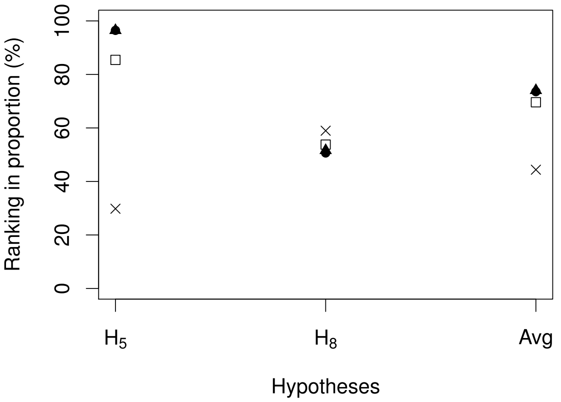

Figure 5: Rankings of unreasonable hypotheses.

Next, we discuss the results for individual hypotheses. The hypotheses concerning Blood Viscosity (i.e., H_2, H_3, and H_4) are, as described, reasonable, and semantic-based reasonability measures, R_{max} and R_{avg} were able to rank them higher than the frequency-based R_{freq}. This is because the frequencies of the events between fish oil and blood viscosity and between blood viscosity and Raynaud's disease were very small; only one and four, respectively. Note that, however, another frequency-based measure R_{tfidf} also worked well for these hypotheses in spite of the low frequencies thanks to the IDF factor which boosts infrequent concepts.

For the remaining reasonable hypotheses, H_1, H_6, and H_7, they were generally ranked lower (having larger values) than the other reasonable hypotheses H_2, H_3, and H_4. This is mainly due to the insufficient number of MeSH terms associated with the events from which the hypotheses were derived. Note that R_{max} worked relatively well even for these difficult hypotheses.

For the unreasonable hypotheses H_5 and H_8, R_{avg} closely followed by R_{max} and R_{tfidf} was able to rank them low (having large values). R_{freq}, on the other hand, performed noticeably worse primarily because their *B*-terms were quite general and highly frequent concepts, resulting in spuriously high similarity values by R_{freq}.

Then, we compare two semantic-based reasonability measures, R_{max} and R_{avg}. While there is no clear difference for the unreasonable hypotheses, R_{max} was generally able to rank the reasonable hypotheses higher than R_{avg}. This observation suggests that R_{max}, which disregards dissimilar concept pairs, is preferred. We present an illustrative example to show why this is the case. Suppose that there are two events, e_a and e_b, each represented by a set of concepts, {Blood Viscosity, Fish Oil} and {Blood Viscosity, Platelet Aggregation}, respectively. According to Equation (1), concept similarities of all combinations of the concepts between the two sets are calculated as follows: sim(Blood Viscosity, Blood Viscosity)=1 (since the term Blood Viscosity has no hyponyms), sim(Blood Viscosity, Platelet Aggregation)=0.64, sim(Fish Oil, Blood Viscosity)=0, and sim(Fish Oil, Platelet Aggregation)=0. In this case, R_{max} and R_{avg}, of the hypotheses derived from e_a and e_b become $(1+0)/2+(1+0.64)/2 = 1.32$ and $(1+0.64+0+0)/2 \cdot 2 = 0.41$, respectively. Further suppose that two events e'_a and e'_b are defined by adding a concept "Vascular Disease" to e_a and "Raynaud Disease" to e_b, respectively. Because these concepts are semantically similar, the event similarity of e'_a and e'_b should not decline much from that of e_a and e_b. The concept similarities involving either "Vascular Disease" or "Raynaud Disease" are all zero except for sim(Raynaud Disease, Vascular Diseases)=0.47. Then, the event similarities between e'_a and e'_b are calculated as R_{max}=$(1+0+0.47)/3+(1+0.64+0.47)/3 \simeq 1.19$ and $R_{avg} = (1+0.64+0+0+0+0+0+0+0.47)/(3 \cdot 3) \simeq 0.234$. While R_{avg} dropped to around the half of that between e_a and e_b, R_{max} decreased only slightly. The undesired behavior of R_{avg} is caused by many zero similarities between dissimilar concepts.

4.3 Additional Experiment

We carried out another experiment using the relation between migraine and magnesium—another hypothesis Swanson discovered [22], so as to examine how our proposed semantic-based reasonability measure works for the different "true" hypothesis.

Similar to the former experiment, we first constructed a biomedical entity network from the literature published between 1966 and 1987 (the relation was discovered in 1988). As a result, we obtained an entity network composed of 29,915 nodes and 260,562 edges after filtering by UMLS semantic types similar to those used by Weeber [25].

Given migraine as the *A*-term, 69,972 hypotheses were obtained, of which there were 90 hypotheses whose *C*-terms were magnesium or magnesium deficiency. We again ranked all the 69,972 hypotheses using the reasonability measures from Section 3.4.

Overall, the results were found similar to the former results regarding Raynaud's disease except that $\mathrm{R_{freq}}$ performed the best for reasonable hypotheses, followed by $\mathrm{R_{max}}$, $\mathrm{R_{tfidf}}$ and $\mathrm{R_{avg}}$. The good performance of $\mathrm{R_{freq}}$ is due to the fact that the frequencies of the events that yielded the reasonable hypotheses were large enough to rank them high by $\mathrm{R_{freq}}$. For unreasonable hypotheses, however, $\mathrm{R_{max}}$ and $\mathrm{R_{avg}}$ worked the best, ranking those hypotheses lower than the other measures.

In summary, the frequency-based measure $\mathrm{R_{freq}}$ could properly rank reasonable hypotheses only if the event frequencies concerning the hypotheses are sufficient, although it generally gives inappropriately high rankings to the hypotheses with general *B*-terms. On the other hand, the semantic-based measure produces relatively stable and proper rankings for both reasonable and unreasonable hypotheses irrespective of event frequencies.

5. CONCLUSION AND FUTURE WORK

In this work, we aimed to identify reasonable hypotheses, especially those derived from low-frequent terms or events—by focusing on the reasonability of the hypotheses. As the first step toward this goal, we assumed that similar events produced a reasonable hypothesis and defined simple event similarities as an extension of concept similarity using the MeSH thesaurus. Using the true hypotheses reported in the hypotheses discovery literature, we conducted comparative experiments, where our semantic-based reasonability measures, $\mathrm{R_{max}}$ and $\mathrm{R_{avg}}$, as well as two frequency-based measures were examined whether they could properly rank reasonable and unreasonbale hypotheses. The results showed that $\mathrm{R_{max}}$ produced stable and appropriate rankings for most cases disregarding the frequencies of the events from which hypotheses were generated. On the other hand, frequency-based measures were by definition directly much influenced by concept/event frequencies and shown not reliable for some cases.

For future work, we will consider the relevance of MeSH terms for their representing events. Also, we plan to exploit UMLS Metathesaurus and WordNet in addition to MeSH for improving the coverage.

6. REFERENCES

[1] Sophia Ananiadou, Douglas B. Kell, and Junichi Tsujii. Text mining and its potential applications in systems biology. *Trends in Biotechnology*, 24(12):571–579, 2006.

[2] Alan R. Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In *Proceedings of American Medical Informatics 2001 Annual Symposium*, pages 17–21, 2001.

[3] Aaron M. Cohen and William R. Hersh. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71, 2005.

[4] Ralph A. DiGiacomo, JM Kremer, and DM Shah. Fish-oil dietary supplementation in patients with Raynaud's phenomenon: A double-blind, controlled, prospective study. *The American Journal of Medicine*, 86(2):158–164, 1989.

[5] E. Ghigo, E. Arvat, G. Rizzi, J. Bellone, M. Nicolosi, G. M. Boffano, M. Mucci, M. F. Boghen, and F. Camanni. Arginine enhances the growth hormone-releasing activity of a synthetic hexapeptide (GHRP-6) in elderly but not in young subjects after oral administration. *Journal of Endocrinol Investigation*, 17:157–162, Mar 1994.

[6] Michael D. Gordon and Robert K. Lindsay. Toward discovery support systems: a replication, re-examination, and extension of Swanson 's work on literature-based discovery of a connection between Raynaud's and fish oil. *Journal of the American Society for Information Science*, 47(2):116–128, 1996.

[7] William R Hersh, Ravi Teja Bhupatiraju, Laura Ross, Phoebe Roberts, Aaron M Cohen, and Dale F Kraemer. Enhancing access to the Bibliome: the TREC 2004 Genomics Track. *Journal of Biomedical Discovery and Collaboration*, 1 2006.

[8] Dimitar Hristovski, Sašo Džeroski, Borut Peterlin, and Anamarija Rožić-Hristovski. Supporting discovery in medicine by association rule mining of bibliographic databases. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 446–451. Springer-Verlag, 2000.

[9] Lars Juhl Jensen, Jasmin Saric, and Peer Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nature reviews genetics*, 7:119–129, 2006.

[10] J. J. Jiang and D. W. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *In International Conference Research on Computational Linguistics*, pages 9008+, September 1997.

[11] Dekang Lin. An information-theoretic definition of similarity. In *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, 1998.

[12] Robert K. Lindsay and Michael D. Gordon. Literature-based discovery by lexical statistics. *Journal of the American Society for Information Science*, 50(7):574–587, 1999.

[13] NLM. Fact sheet medical subject headings, 2008.

[14] N. M. Ramadan, H. Halvorson, A. Vande-Linde, S. R. Levine, J. A. Helpern, and K. M. Welch. Low brain magnesium in migraine. *Headache*, 29:416–419, 1989.

[15] P. Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, November 1995.

[16] Jesse A. Stumpc Ronald N. Kostoffa, Joel A. Blockb and Dustin Johnson. Literature-related discovery (lrd): Potential treatments for raynaud's phenomenon. *Technological Forecasting and Social Change*,

75(2):203–214, 2008.

[17] Nuno Seco, Tony Veale, and Jer Hayes. An intrinsic information content metric for semantic similarity in wordnet. In *Proceedings of European Conference on Artificial Intelligence 2004*, pages 1089–1090, 2004.

[18] Karen Sparck Jones. Statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–20, 1972.

[19] Padmini Srinivasan. Text mining: generating hypotheses from medline. *Journal of the American Society for Information Science and Technology*, 55(5):396–413, 2004.

[20] D. R. Swanson. Somatomedin C and arginine: implicit connections between mutually isolated literatures. *Perspect. Biol. Med.*, 33:157–186, 1990.

[21] Don R. Swanson. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18, 1986.

[22] Don R. Swanson. Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, 31(4):526–557, 1988.

[23] Don R. Swanson and Neil R. Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91(2):183–203, 1997.

[24] Pratt Wanda and Yetisgen-Yildiz Meliha. Litlinker: capturing connections across the biomedical literatur. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 105–112, 2003.

[25] Marc Weeber, Henry Klein, Lolkje T. W. de Jong-van den Berg, and Rein Vos. 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.

[26] Marc Weeber, Rein Vos, Henny Klein, Lolkje T.W. de Jong-van den Berg, Alan R. Aronson, and Grietje Molema. Generating Hypotheses by Discovering Implicit Associations in the Literature: A Case Report of a Search for New Potential Therapeutic Uses for Thalidomide. *Journal of the American Medical Informatics Association*, 10(3):252–259, 2003.

[27] Jonathan Wren. Extending the mutual information measure to rank inferred literature relationships. *BMC Bioinformatics*, 5(1), 2004.