IPSJ Transactions on Bioinformatics Vol. 4 9–20 (May 2011)

Original Paper

Hypothesis Ranking Based on Semantic Event Similarities

Taiki Miyanishi,[†1] Kazuhiro Seki[†1]
and Kuniaki Uehara[†1]

Accelerated by the technological advances in the biomedical domain, the size of its literature has been growing very rapidly. As a consequence, 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. We developed a prototype system called, *Hypothesis Explorer*, and conducted evaluative experiments through which the validity of our approach is demonstrated in comparison with those based on term frequencies, often adopted in the previous work.

1. Introduction

The biomedical literature has been rapidly growing at a rate no one can keep up with, which makes it infeasible for individual researchers to comprehend all the information related to their interests [1)–3)]. For this reason, it is conceivable that much potential knowledge, or *hypotheses*, remains 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)–9)].

As the pioneering work for biomedical hypotheses discovery, Swanson [8)] 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 [10)–21)]. 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 despite 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 the present paper. Section 3 details our developed framework for hypothesis generation and ranking based on event similarities. Section 4 introduces our hypothesis generation system, *Hypothesis Explorer*, which helps us find new hypotheses by visualizing explicit/implicit relationships between concepts and their supporting information. Section 5 evaluates our hypothesis ranking functions in comparison with term frequency-based functions. Finally, Section 6 concludes with a 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 association 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 an explicit relationship reported in the literature. This one directional exploration starting from only one concept, A, for hypothesis discovery is called "open" discovery. Hypothesis discovery can also be

†1 Graduate School of System Informatics, Kobe University

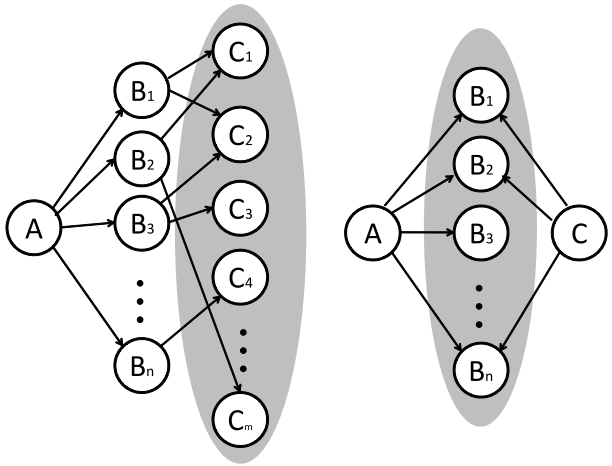


Fig. 1 Open and closed discovery.

accomplished bidirectionally given two concepts, A and C, which focuses more on identifying intermediate concepts that connects the two starting concepts and is called "closed discovery". The left figure in **Fig. 1** provides an example of open discovery which attempts to find indirect, undiscovered relationships between a given entity (A) and other entities (C) where the C terms can be any concepts. On the other hand, the right figure in Fig. 1 is an example of closed discovery which aims to find intermediate concepts (B) that provide logical connections to the given two concepts, A and C. Note that in this case A and C may have a known association but their connections are unknown.

For open discovery, it may sound trivial to predict the potential relationship 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 literature representing two different specialties, such as nutrition and genetics, 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 [22),23)], have developed computer systems to aid hypothesis discovery [10)–21)]. Some of them support only closed discovery [19),21)–23)], while our system presented in Section 4 can be applied to both open and closed discovery. In the following, we focus on two important attempts most related to the present work.

Weeber [15)] 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 [22)], each sentence in textual portions of MEDLINE records was mapped to the concepts defined in the UMLS Metathesaurus by the MetaMap program [24)] 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." [⋆2] 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 [14)] developed another system, called Manjal [⋆3], 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

⋆1 UMLS is an NLM's project to develop and distribute multi-purpose, electronic knowledge sources and its associated lexical programs. http://www.nlm.nih.gov/research/umls/
⋆2 There are 134 semantic types in total.
⋆3 http://sulu.info-science.uiowa.edu/Manjal.html

into the corresponding UMLS semantic types according to a predefined mapping. Similar to the approach by Weeber [15], 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 [25] 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 numerous hypotheses generated. 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 criterion, called *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 [15]. Weeber used MetaMap program [24] to obtain UMLS concepts and extracted relationships between concepts based on their co-occurrences 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 a concise, high quality summary of the articles as reported in the functional genomics domain [26]. In addition, it is expected that resulting hypotheses become more precise by limiting our focus to titles and MeSH as reported by Swanson, et al [23]. 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. For example, if two relationships between A and B and between B and C are extracted simply based on co-occurrences from two sentences, "A causes B" and "B does *not* cause C", it would result in a false hypothesis "A (causes) C". In the present work, however, we use co-occurrences for simplicity and attempt to minimize the risk to produce such false hypotheses by looking at only titles which were found often affirmative by our informal observation. 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 B concepts and for terminal C concepts, respectively.

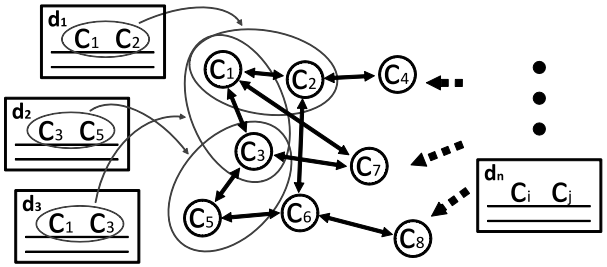


Fig. 2 Entity network constructed from the literature.

We call a co-occurrence of two concepts in an article title an *event* disregarding the type of the event. Then, by merging common concepts of the extracted events, an entity network can be constructed. **Figure 2** 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 Fig. 2, event c_1–c_2 (composed of the two 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 should be emphasized that it is only possible to discover these indirect relationships by gathering the information found in separate articles, d_2 and d_3, together.

More formally, hypothesis generation can be performed by exploring the entity

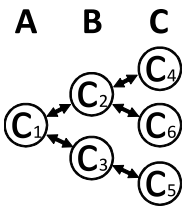


Fig. 3 Example of A, B, and C terms.

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. In this paper, we limit our discussion to a depth of two for simplicity. The discovered paths between starting and ending points are interpreted as hypotheses except for closed chains in the network. As a matter of convenience, we call the starting, ending, and intermediate concepts A term, C term, and B term, respectively. **Figure 3** shows the resulting pathways aligned by A, B, and C terms based on Fig. 2, where node c_1 was used as a starting point for search.

Although it is not the focus of the present work, there is still room for further investigation in our approach to hypothesis generation. For example, A, B, C terms are concepts defined in the UMLS Metathesaurus, where each concept is categorized under semantic hierarchies. Considering such hierarchies would allow us to associate related concepts in the network and consequently yield more hypotheses. The use of the thesaurus in hypothesis generation should be studied in future work.

3.2 Reasonability of Hypothesis

The generated hypothesis becomes new knowledge only after verifying it through actual experiments, which are often costly—or infeasible—to carry out. Our purpose is to identify reasonable hypotheses that are more likely to lead to new knowledge among automatically derived 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

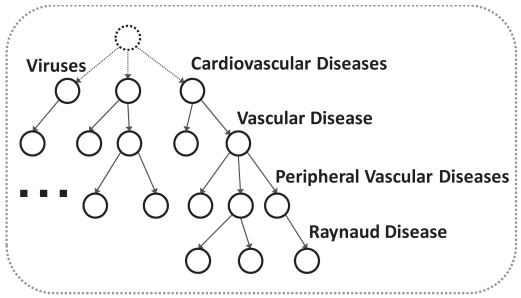


Fig. 4 Fragment of the MeSH thesaurus.

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.

Now, 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 by hand approximately ten MeSH terms on average. The 2010 version of MeSH contains a total of 25,588 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 [27]. MeSH terms can be suitable features to represent an article due to high quality assignment by experts. **Figure 4** presents a fragment of the MeSH thesaurus.

Here, a potential problem regarding MeSH terms is that they characterize not an event but (some parts of) 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 hopefully MeSH terms well represent.

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 [28)–31)] where each concept is categorized under semantic hierarchies.

Among them, we adopt the measure proposed by Seco [31] 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, and defined the similarity between two concepts according to the information value of their common ancestor, which depends on its position in the thesaurus. 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

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

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

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, 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 Eq. (2), which is equivalent to the value of the least informative concept, serves as a normalizing factor to ensure that IC values are in the range from 0 to 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

This section extends the concept similarity defined in Eq. (1) 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 **Fig. 5**. 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.

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

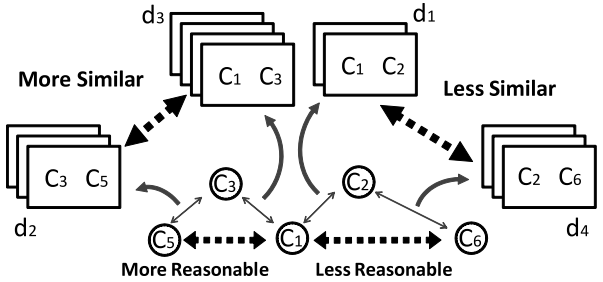


Fig. 5 Relation between event similarity and hypothesis reasonability.

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

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 extracting MeSH terms from the MEDLINE record in which an event is found. When the same event (co-occurrence of two concepts) is found in multiple articles, the MeSH terms assigned to those articles are aggregated into $\mathbf{M}$. A shortcoming of this similarity $\mathrm{R_{avg}}$ is that it considers the similarities even between dissimilar concepts as we will discuss in Section 5.

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.

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

$$+ \frac{1}{|\mathbf{M}_j|} \sum_{m_l \in \mathbf{M}_j} \max_{m_k \in \mathbf{M}_i} \mathrm{sim}(m_k, m_l) \quad (4)$$

3.4.3 Event Similarity by TFIDF

To compare with the above two semantic-based ranking functions, we also introduce a frequency-based event similarity adopting the often-used TFIDF term weighting scheme. This definition resembles the one proposed by Srinivasan [14].

As mentioned above, an event may be extracted from multiple articles, and thus, it is associated with duplicated MeSH terms when those articles have the same MeSH terms. We regard the number of duplicates of a MeSH term m_j for event e_i as its term frequency (denoted as n_{ij}) and the number of articles indexed with the MeSH term as its document frequency (denoted as n_j). Then, each event e_i can be represented as a MeSH term vector weighted by TFIDF. That is,

$$\boldsymbol{e}_i = (w_{i1}, w_{i2}, \cdots, w_{iM}) \quad (5)$$

where w_{ij} is the TFIDF value, defined as $n_{ij} \times \log(N/n_j)$ for MeSH term m_j. M denotes the total number of MeSH terms, and N denotes the total number of documents. Using the MeSH vectors, the similarity between events e_i and e_j can be computed by cosine similarity as shown in Eq. (6) as normally done in the information retrieval literature.

$$\mathrm{R}_{\mathrm{tfidf}}(e_i, e_j) = \frac{\boldsymbol{e}_i \cdot \boldsymbol{e}_j}{|\boldsymbol{e}_i||\boldsymbol{e}_j|} \quad (6)$$

3.4.4 Event Frequency

We define another event reasonability measure $\mathrm{R}_{\mathrm{freq}}$ based not on event similarity but on event frequencies as follows:

$$\mathrm{R}_{\mathrm{freq}}(e_i, e_j) = \sqrt{freq(e_i) \times freq(e_j)} \quad (7)$$

where $freq(e)$ denotes the frequency of an event e. The intuition behind this is that a hypothesis supported by highly frequent events should be reasonable. By comparing $\mathrm{R}_{\mathrm{freq}}$ with other measures based on event similarity, we can contrast the difference between frequency- and similarity-based measures.

4. Hypothesis Explorer

Based on the framework described in Section 3, we implemented a prototype of a hypothesis discovery system, called *Hypothesis Explorer*, to allow us to efficiently find new hypotheses. Given a starting term (A term), and optionally a terminal term (C term), *Hypothesis Explorer* visualizes a biomedical entity network extracted from the biomedical literature and automatically discovers explicit associations (i.e., not directly linked entities) as hypotheses. It can also provide supporting evidence (article titles) of events for review from which the hypotheses were derived. **Figure 6** presents an example of closed discovery when

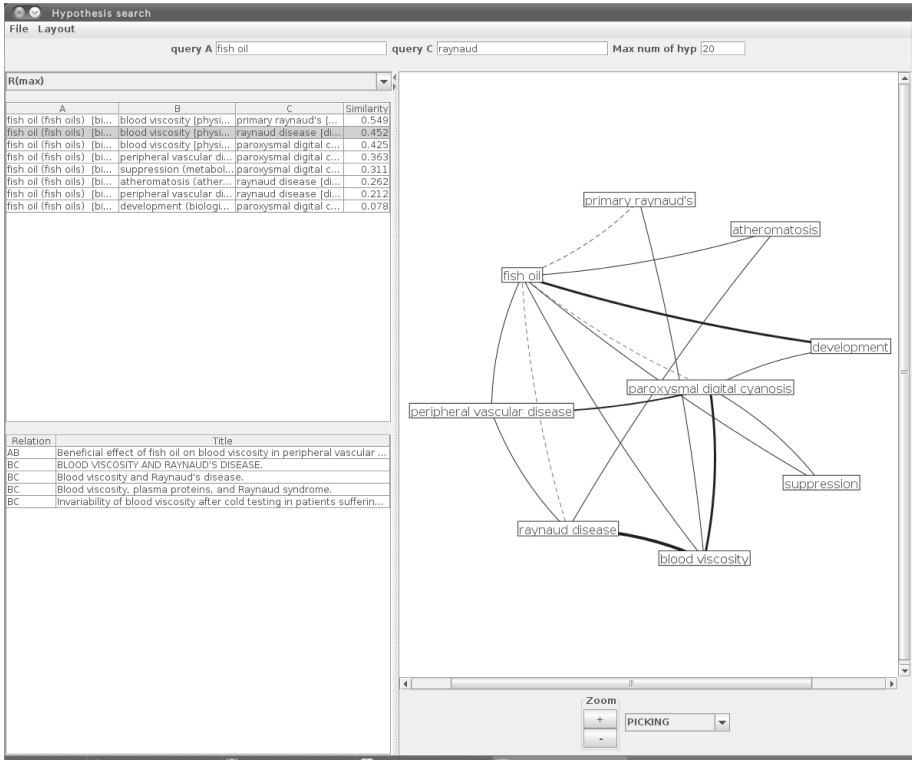


Fig. 6 Screen shot of *Hypothesis Explorer*.

"fish oil" and "raynaud" were used as A term and C term, respectively.

This system mainly consists of three regions, three text boxes, and one dropdown menus. The top text boxes accept A term and a C term as user queries. To run open discovery, we feed only A term and for closed discovery we feed both A and C terms. In addition, the rightmost text box in the top can be used to specify the maximum number of hypotheses displayed on the windows to avoid a clutter as there are often a large number of hypotheses generated.

The top left region lists generated hypotheses composed of sets of A, B, and C terms in the descending order of their semantic similarities shown in the rightmost column. The bottom left region shows the source article titles from which a hypothesis is generated. Remember that a hypothesis is made up of two events, and each event is extracted from article titles as a co-occurrence of two concepts (i.e., A and B terms or B and C terms). When a hypothesis is clicked in the top left region, it is highlighted and the article titles containing the A-B and B-C events for the hypothesis are displayed the bottom left region. This functionality allows the users of the system to easily assess the validity of the events and the hypothesis. The right big region visualizes the biomedical entity network constructed from the literature, where generated hypotheses are displayed as red dotted lines which connect A and C terms. On the other hand, the black solid lines indicate explicit relationships or events which are part of the generated hypotheses. The thickness of the black lines corresponds to the frequencies of the respective events. The concept in blue font indicates the B term for the hypothesis selected in the top left region.

Lastly, the dropdown menu above the top left region allows us to specify which similarity measure we use to explore hypotheses. For the example in Fig. 6, R_{max} is selected as event similarity (i.e., reasonability).

5. Evaluation

5.1 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 in comparison with other measures discussed in Section 3.4.

Biologic Function, Cell Function, Disease or Syndrome, Lipid, Molecular Function, Organ or Tissue Function, Organism Function, Pathologic Function, Physiologic Function

Fig. 7 UMLS semantic types.

To be precise, we looked at the biomedical literature published between 1960 and 1985, and used their titles for constructing an entity network as described in Section 3.1. 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.

Figure 7 shows nine UMLS semantic types, which were also used by Weeber [15], to restrict hypothesis exploration in the relationship extraction step. 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.

5.2 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 as an open discovery. Among them, there were eight hypotheses whose C term was Raynaund's disease or its synonyms. **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

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

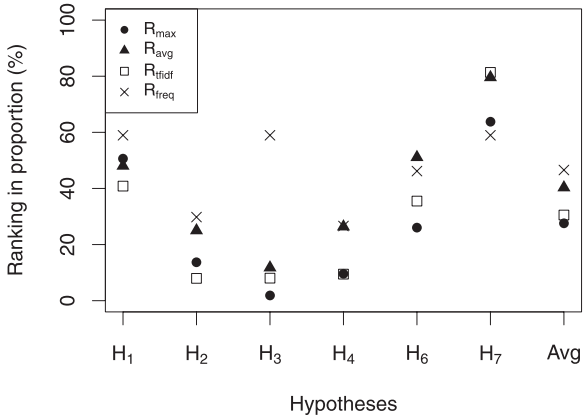


Fig. 8 Rankings of reasonable hypotheses.

"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 8** and **9** 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 Fig. 8 and Fig. 9 are the respective average rankings

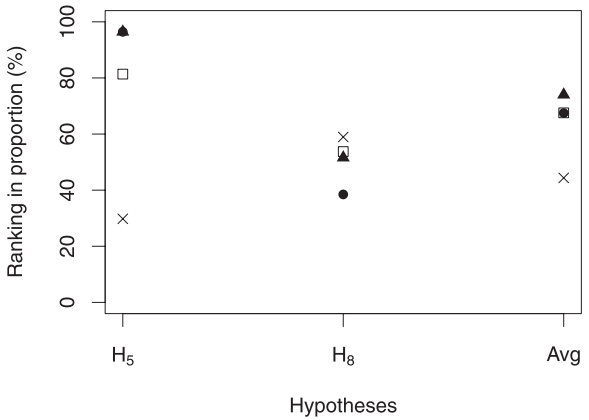


Fig. 9 Rankings of unreasonable hypotheses.

obtained by different reasonability measures. To remind, R_{avg} is the reasonability measure defined as the average of the similarities for all 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 Fig. 8 are considered better when they have smaller 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, Fig. 9 shows that R_{avg} was able to rank them lower (i.e., higher values) than the other reasonability measures. R_{max} and R_{tfidf} worked comparably.

Next, we discuss the results for individual hypotheses. The hypotheses concerning Blood Viscosity (i.e., H_2, H_3, and H_4) are judged reasonable as mentioned, and semantic-based reasonability measures, R_{max} and R_{avg} were generally 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 hypothesis H_5, R_{avg} and R_{max} were able to rank it lower (having larger values), followed by R_{tfidf}. 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}. As for H_8, it was found that R_{max} did not work well as compared to the other reasonabilities. We will discuss the property of R_{max} in comparison with R_{avg} shortly.

Looking at two semantic-based reasonability measures, R_{max} and R_{avg}, the former works better than the latter except for a couple of instances, including H_8. This observation suggests that R_{max}, which disregards dissimilar concept pairs, is preferred over R_{avg}. 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 Eq. (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. On the contrary, R_{max} focuses only on similar concepts and is free from such problems.

Acquired Abnormality; Amino Acid, Peptide, or Protein; Anatomical Abnormality; Antibiotic; Biologically Active Substance; Biomedical or Dental Material; Body Substance; Carbohydrate; Cell Function; Congenital Abnormality; Disease or Syndrome; Eicosanoid; Element, Ion, or Isotope; Enzyme; Finding; Food; Hazardous or Poisonous Substance; Health Care Related Organization; Hormone; Immunologic Factor; Indicator, Reagent, or Diagnostic Aid; Injury or Poisoning; Inorganic Chemical; Intellectual Product; Invertebrate; Laboratory Procedure; Lipid; Manufactured Object; Molecular Function; Neuroreactive Substance or Biogenic Amine; Organ or Tissue Function; Organic Chemical; Organism Function; Organophosphorus Compound; Pathologic Function; Pharmacologic Substance; Physiologic Function; Sign or Symptom; Steroid; Vitamin

Fig. 10 UMLS semantic types used for migraine and magnesium.

5.3 Additional Experiment

We carried out another experiment using a known relation between migraine and magnesium—another hypothesis Swanson discovered [9], so as to examine how our proposed semantic-based reasonability measure works for a different "true" hypothesis.

Similar to the former experiment, we first constructed a biomedical entity network from the literature published between 1966 and 1987 (since 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 [15] as shown in **Fig. 10**.

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. Because there are many hypotheses generated for this experiment, we only present the average rankings (in proportion to the total number of hypothe-

Table 2 Average rankings in proportion (%) for the case of migraine and magnesium.

	R_{max}	R_{avg}	R_{tfidf}	R_{freq}
Reasonable	36.6	47.6	36.3	22.6
Unreasonable	53.4	54.1	42.0	45.9

ses) for reasonable and unreasonable hypotheses in **Table 2**. Overall, the results are similar to the previous results regarding Raynaud's disease except that R_{freq} performed the best for reasonable hypotheses, followed by R_{max} and R_{tfidf}. The good performance of 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 R_{freq}. For unreasonable hypotheses, however, R_{max} and R_{avg} worked the best, ranking those hypotheses lower than the other measures.

In summary, the frequency-based measure 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.

6. Conclusion and Future Work

In this work, we aimed to identify reasonable hypotheses, especially those derived from infrequent 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. We developed a prototype hypothesis discovery system, *Hypothesis Explorer*, implementing our proposed framework that supports hypothesis discovery through automatic hypothesis generation and ranking and biomedical concept network visualization for a given starting concept (and a terminal concept). Using the true hypotheses reported in the hypotheses discovery literature, we conducted comparative experiments, where our semantic-based reasonability measures, R_{max} and R_{avg}, as well as two frequency-based measures were examined whether they could properly rank reasonable and unreasonable hypotheses. The results showed that 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 respective events. Currently, we indiscriminately use all the MeSH terms assigned to the articles from which an event is extracted for representing the event. It is not likely that those MeSH terms are all relevant to the event and need to be distinguished according to their relevance. Another issue is the coverage of hypothesis discovery. Our analysis is limited to article titles and also does not consider semantic hierarchy of concepts in hypothesis generation. To deal with it, we plan to exploit UMLS Metathesaurus and WordNet in addition to MeSH, and will try to extract biomedical relationships from not only titles but also abstracts which have more complete information available for hypotheses generation. Lastly, we would like to collaborate with biomedical experts to validate unknown hypotheses generated from our proposed framework in direct comparison with the others.

Acknowledgments We would like to thank Mathieu Blondel for implementing the prototype *Hypothesis Explorer* system. This work was supported by KAKENHI #21700169 and Hyogo Science and Technology Association Grant #22S003.

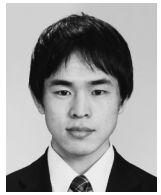
References

1) Ananiadou, S., Kell, D.B. and Tsujii, J.: Text mining and its potential applications in systems biology, *Trends in Biotechnology*, Vol.24, No.12, pp.571–579 (2006).
2) Cohen, A.M. and Hersh, W.R.: A survey of current work in biomedical text mining, *Briefings in Bioinformatics*, Vol.6, No.1, pp.57–71 (2005).
3) Jensen, L.J., Saric, J. and Bork, P.: Literature mining for the biologist: From information retrieval to biological discovery, *Nature Reviews Genetics*, Vol.7, pp.119–129 (2006).
4) DiGiacomo, R.A., Kremer, J. and Shah, D.: Fish-oil Dietary Supplementation in Patients with Raynaud's Phenomenon: A Double-Blind, Controlled, Prospective Study, *The American Journal of Medicine*, Vol.86, No.2, pp.158–164 (1989).
5) Ghigo, E., Arvat, E., Rizzi, G., Bellone, J., Nicolosi, M., Boffano, G.M., Mucci, M., Boghen, M.F. and Camanni, F.: Arginine enhances the growth hormone-releasing activity of a synthetic hexapeptide (GHRP-6) in elderly but not in young subjects

after oral administration, *J. Endocrinol Inv.*, Vol.17, pp.157–162 (1994).
6) Ramadan, N.M., Halvorson, H., Vande-Linde, A., Levine, S.R., Helpern, J.A. and Welch, K.M.: Low brain magnesium in migraine, *Headache*, Vol.29, pp.416–419 (1989).
7) Swanson, D.R.: Somatomedin C and arginine: Implicit connections between mutually isolated literatures, *Perspect. Biol. Med.*, Vol.33, pp.157–186 (1990).
8) Swanson, D.R.: Fish oil, Raynaud's syndrome, and undiscovered public knowledge, *Perspect. Biol. Med.*, Vol.30, No.1, pp.7–18 (1986).
9) Swanson, D.R.: Migraine and magnesium: Eleven neglected connections, *Perspect. Biol. Med.*, Vol.31, No.4, pp.526–557 (1988).
10) Gordon, M.D. and Lindsay, R.K.: 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, *J. Am. Soc. Inf. Sci.*, Vol.47, No.2, pp.116–128 (online), DOI:http://dx.doi.org/10.1002/(SICI)1097-4571(199602)47:2⟨*116::AID-ASI3*⟩3.3.CO;2-P (1996).
11) Hristovski, D., Džeroski, S., Peterlin, B. and Rožić-Hristovski, A.: Supporting Discovery in Medicine by Association Rule Mining of Bibliographic Databases, *Proc. 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pp.446–451, Springer-Verlag (2000).
12) Wanda, P. and Meliha, Y.-Y.: LitLinker: Capturing connections across the biomedical literatur, *Proc. 2nd international conference on Knowledge capture*, pp.105–112 (2003).
13) Lindsay, R.K. and Gordon, M.D.: Literature-based discovery by lexical statistics, *J. Am. Soc. Inf. Sci.*, Vol.50, No.7, pp.574–587 (online), DOI:http://dx.doi.org/10.1002/(SICI)1097-4571(1999)50:7⟨*574::AID-ASI3*⟩3.0.CO;2-Q (1999).
14) Srinivasan, P.: Text mining: Generating hypotheses from MEDLINE, *J. Am. Soc. Inf. Sci. Technol.*, Vol.55, No.5, pp.396–413 (online), DOI:http://dx.doi.org/10.1002/asi.10389 (2004).
15) Weeber, M., Klein, H., de Jong-van den Berg, L.T.W. and Vos, R.: Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries, *J. Am. Soc. Inf. Sci. Technol.*, Vol.52, No.7, pp.548–557 (2001).
16) Weeber, M., Vos, R., Klein, H., de Jong-van den Berg, L.T.W., Aronson, A.R. and Molema, G.: Generating Hypotheses by Discovering Implicit Associations in the Literature: A Case Report of a Search for New Potential Therapeutic Uses for Thalidomide, *J. Am. Med. Inf. Assoc.*, Vol.10, No.3, pp.252–259 (online), available from ⟨http://www.jamia.org/cgi/content/abstract/10/3/252⟩ (2003).
17) Wren, J.: Extending the mutual information measure to rank inferred literature relationships, *BMC Bioinformatics*, Vol.5, No.1, p.145 (2004).
18) Kostoff, R.N., Block, J.A., Stumpc, J.A. and Johnson, D.: Literature-related discovery (LRD): Potential treatments for Raynaud's Phenomenon, *Technological Forecasting and Social Change*, Vol.75, No.2, pp.203–214 (2008).
19) Smalheiser, N.R., Torvika, V.I. and Wei, Z.: Arrowsmith two-node search interface: A tutorial on finding meaningful links between two disparate sets of articles in MEDLINE, *Computer Methods and Programs in Biomedicine*, Vol.94, No.2, pp.190–197 (2009).
20) Trevor, C., Roger, S. and Widdows, D.: Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections, *J. Biomed. Inf.*, Vol.43, No.2, pp.240–256 (2010).
21) Torvik, V. and Smalheiser, N.: A quantitative model for linking two disparate sets of articles in Medline, *Bioinformatics*, Vol.23, No.13, pp.1658–1665 (2007).
22) Swanson, D.R. and Smalheiser, N.R.: An interactive system for finding complementary literatures: A stimulus to scientific discovery, *Artif. Intell.*, Vol.91, No.2, pp.183–203 (online), DOI:http://dx.doi.org/10.1016/S0004-3702(97)00008-8 (1997).
23) Swanson, D.R., Smalheiser, N.R. and Torvik, V.I.: Ranking indirect connections in literature-based discovery: The role of medical subject headings, *J. Am. Soc. Inf. Sci. Technol.*, Vol.57, No.11, pp.1427–1439 (2006).
24) Aronson, A.R.: Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program, *Proc. American Medical Informatics 2001 Annual Symposium*, pp.17–21 (2001).
25) Sparck Jones, K.: Statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation*, Vol.28, No.1, pp.11–20 (1972).
26) Hersh, W.R., Bhupatiraju, R.T., Ross, L., Roberts, P., Cohen, A.M. and Kraemer, D.F.: Enhancing access to the Bibliome: The TREC 2004 Genomics Track, *Journal of Biomedical Discovery and Collaboration*, Vol.1, p.3 (2006).
27) NLM: Fact Sheet Medical Subject Headings (2008).
28) Jiang, J.J. and Conrath, D.W.: Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, *International Conference Research on Computational Linguistics* (1997).
29) Lin, D.: An Information-Theoretic Definition of Similarity, *Proc. 15th International Conference on Machine Learning*, pp.296–304 (1998).
30) Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy, *Proc. 14th International Joint Conference on Artificial Intelligence*, pp.448–453 (1995).
31) Seco, N., Veale, T. and Hayes, J.: An Intrinsic Information Content Metric for Semantic Similarity in WordNet, *Proc. European Conference on Artificial Intelligence 2004*, pp.1089–1090 (2004).

(Received December 10, 2010)
(Accepted February 14, 2011)
(Released May 18, 2011)

(Communicated by *Tetsuo Shibuya*)



Taiki Miyanishi was born in 1987. He received his M.S. in engineering from Kobe University. His research interests are machine learning, link mining, and natural language processing. He is currently a Ph.D. student in the Graduate School of System Infomatics, Kobe University. He is a member of the Japanese Society for Artificial Intelligence.

Kazuhiro Seki received his Ph.D. in information science from Indiana University, Bloomington. His research interests are in the areas of natural language processing, information retrieval, machine learning, and their applications to intelligent information processing and management systems. He is currently an associate professor in the Graduate School of System Informatics at Kobe University, Japan.

Kuniaki Uehara received his B.E., M.E. and D.E. degrees in information and computer sciences from Osaka University, Japan. He was an assistant professor in the Institute of Scientific and Industrial Research at Osaka University and was a visiting assistant professor at Oregon State University. Currently, he is a professor in the Graduate School of System Informatics at Kobe University, Japan. His conducting research is in the areas of machine learning, data mining, and multimedia processing. He is a member of the Information Processing Society of Japan, Japan Society for Software Science and Technology, and AAAI.