Claremont Colleges
Scholarship @ Claremont

CGU Faculty Publications and Research | CGU Faculty Scholarship

1-1-2008

# The Impact of Directionality in Predications on Text Mining

Gondy A. Leroy
*Claremont Graduate University*

Marcelo Fiszman

Thomas C. Rindflesch

Recommended Citation
Leroy, G.; Fiszman, M.; Rindflesch, T.C.. "The impact of directionality in predications on text mining" 2008 41st Annual Hawaii International Conference on System Sciences. (2008):1826-1835

This Article is brought to you for free and open access by the CGU Faculty Scholarship at Scholarship @ Claremont. It has been accepted for inclusion in CGU Faculty Publications and Research by an authorized administrator of Scholarship @ Claremont. For more information, please contact scholarship@cuc.claremont.edu.

# The Impact of Directionality in Predications on Text Mining

*Gondy Leroy[1], Marcelo Fiszman[2], Thomas C. Rindflesch[2]*
*[1]School of Information Systems and Technology, Claremont Graduate University; [2]Lister Hill National Center for Biomedical Communications, National Library of Medicine*

## Abstract

*The number of publications in biomedicine is increasing enormously each year. To help researchers digest the information in these documents, text mining tools are being developed that present co-occurrence relations between concepts. Statistical measures are used to mine interesting subsets of relations. We demonstrate how directionality of these relations affects interestingness. Support and confidence, simple data mining statistics, are used as proxies for interestingness metrics. We first built a test bed of 126,404 directional relations extracted from biomedical abstracts, which we represent as graphs containing a central starting concept and 2 rings of associated relations. We manipulated directionality in four ways and randomly selected 100 starting concepts as a test sample for each graph type. Finally, we calculated the number of relations and their support and confidence. Variation in directionality significantly affected the number of relations as well as the support and confidence of the four graph types.*

## 1. Introduction

Every year, the number of publications available to researchers increases. This is especially true in biomedicine and genomics. This growth in publications is attributable to many factors, including faster data gathering and processing and faster publication cycles. Although more information may lead to more discoveries and knowledge, for that to happen the available information needs to be read and understood by researchers.

This burgeoning information makes it progressively more difficult to stay up to date on any particular topic. To demonstrate the problem, we used simple keywords to search for five topics in three online databases. Figure 1 shows the increasing number of publications for each year. We searched PubMed (www.pubmed.gov) for 'p53', 'brca1' and 'autism', IEEE Xplore (http://ieeexplore.ieee.org) for 'genetic algorithm', and PsycInfo (http://psycinfo2.apa.org/) for 'depression'. We performed all 5 searches for each year from 1970 to 2006. The resulting graph shows the considerable, in some cases exponential, growth in the number of publications that become available each year—too many for any individual researcher to read and digest.

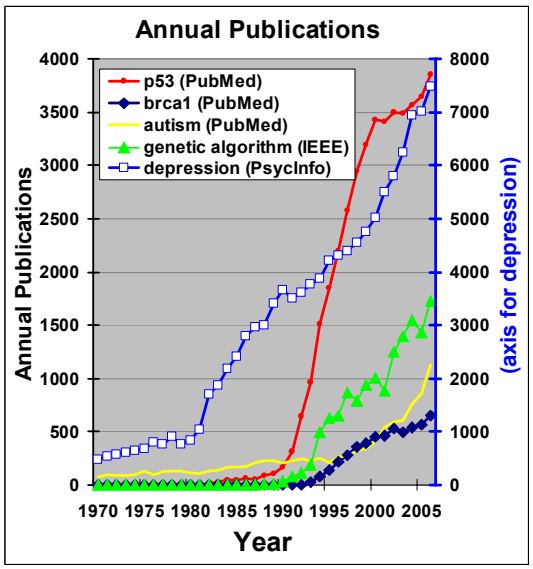


**Figure 1: Examples of increasing number of publications for different topics in different databases (searched on 5/2007)**

In response to this large amount of information, text mining tools are being developed. Fan [4] provides an overview of commercial, generic tools that are available. The goal of such tools is to provide more effective access to information in unstructured text. In biomedicine, two approaches are common. The first treats words and phrases as data elements and calculates their co-occurrence in text. The outcome is often displayed as a visual graph for researchers to browse. The second approach extracts very specific information in the form of



Authorized licensed use limited to: to the Claremont Colleges!. Downloaded on January 22, 2010 at 12:25 from IEEE Xplore. Restrictions apply.

entities from text, for example words or phrases describing genes or proteins. Similar to the first approach, these co-occurring entities are then visualized. Ultimately, this text-based data will be integrated with gene expression data. Currently, most genomics-related text mining work uses data mining and statistical techniques to limit the set of relations between terms that need to be visualized.

Our goal is to show how natural language processing (NLP) that goes beyond phrase and entity extraction and assigns directionality to relations between terms fundamentally affects the existing metrics used to select co-occurrence based relations. We developed a test bed to demonstrate our approach and used NLP techniques to extract terms and the relations between them from text. We use 'term' to refer to both single- and multiple-word noun phrases. We evaluate small graphs that represent samples of the visual displays commonly used with text mining. Three types of directionality are compared against a baseline (no-directionality). It is not our goal to find the most interesting set of relations. Our goal is to show the impact of directionality in graphs on automated and objective measures of interestingness. Therefore, we generate different types of graphs and evaluate them with simple and well known statistics: support and confidence. More complicated measures could be used, but these are often variants of support and confidence and we prefer to use simple measure to facilitate focus on directionality.

## 2. Text Mining in Biomedicine

Researchers have mined text for interesting associations for many years. This idea was always implicitly present when reading literature about a topic. The first one to exploit this approach for finding associations in different sub-domains was Swanson [24]. He devised a method for discovering a relationship between two terms A and C not explicitly mentioned in the literature by uncovering a third, intermediate, term B. See Figure 2 for an overview. Swanson, along with several subsequent systems (e.g. [6, 8, 23, 26]), relied on co-occurrence of A, B, and C terms in corresponding A, B, and C literature domains. In Swanson's original work [2], the co-occurrence of fish oil (A) and blood viscosity (B) on the one hand, and the co-occurrence of blood viscosity and Raynaud's disease (C) on the other, supported his suggestion of fish oil as a treatment for Raynaud's disease.

Today, natural language processing is often combined with statistical calculations. Ideally, the process leading to discoveries such as those made by Swanson could be automated. Unfortunately, current approaches are not sufficiently advanced. In some cases, terms or phrases representing biological entities are extracted from text and co-occurrence is used to establish an association between terms (e.g. [27]). Other systems use more extensive NLP to identify and extract specific relations between terms. Some focus on particular relations, such as those involved in protein interactions (e.g. [3, 13, 25], while other systems accommodate a wider range of relations. For example, Genescene [14, 15] uses a shallow parser to extract a variety of relations between terms. SemRep [20, 21] (used for this work) uses underspecified syntactic analysis and biomedical domain knowledge from the Unified Medical Language System (UMLS). BioMedLee [16] relies on a locally developed semantic lexicon, a grammar formalism that combines syntax and semantics, and a frame-based representation.

Some systems visually display extracted relations in a graph, (e.g. [14, 15, 27]), with terms represented as nodes and the relations between them as edges. Few biomedical NLP tools are offered to the research community for usage. Exceptions are SemRep (http://skr.nlm.nih.gov), BioMedLee (http://zellig.cpmc.columbia.edu/medlee), and GATE, a toolkit that can accommodate NLP development for biomedical text (http://gate.ac.uk/).

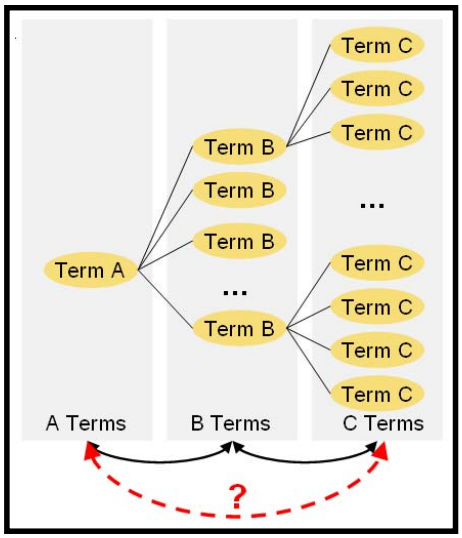


**Figure 2: visual representation of an early, manual approach to text mining**



Authorized licensed use limited to: to the Claremont Colleges!. Downloaded on January 22, 2010 at 12:25 from IEEE Xplore. Restrictions apply.

## 2.1 Term Selection

Terms can be selected in a variety of ways. The best approach depends on the application. For example, to get a broad overview it may be better to include many different types of terms, but to find functional genetic pathways, it is better to focus on genes and proteins. The approach chosen will affect how many terms are available for constructing relations.

The easiest and fastest way to extract terms from text is to accept any noun or noun phrase. A stop word approach (stop words or other symbols as phrase delimiters) or statistical approaches can be used to detect noun phrases quickly. The use of grammatical rules to define noun phrases is a more precise but more demanding method. The tools mentioned above, SemRep and BioMedlEE, are rule-based approaches leveraging grammar rules in combination with lexicons. GATE provides developers the opportunity to develop rules with its jape files and to include external knowledge sources.

In many cases, however, too many terms are extracted per document, and *interesting* or *good* terms then need to be selected from the entire set. The most popular, statistical approaches are tf-idf measures or variants thereof. The tf-idf model calculates how frequently a term appears in a document (tf or term frequency) versus in the document collection (idf or inverse document frequency) to indicate interestingness. Usually, terms that appear frequently in one document but not in every document of the collection are considered more interesting. For example, Jayadevaprakash et al. [11] used tf-idf as their statistical model to extract terms. In addition to statistical approaches, lexicons or ontologies are also used to identify subsets of terms for further use.

## 2.2 Relation Selection

In biomedicine, relations between terms are often visualized as two dimensional graphs. Regardless of the term selection method used, the number of available associations is usually too large to be shown in one graph. A possible solution is to retain all terms and relations but allow users to zoom in and out, making the graph readable when focused.

Another, more common, approach is to select a subset of relations that should be visualized for the user. Several formulas can then be employed to eliminate unwanted relations. Biomedical applications have often used straightforward co-occurrence measures. However, many statistical measures are available that each provide a slightly different focus to select a subset of relations that will be interesting to users. Such interestingness measures have been defined and tested in data mining. Geng and Hamilton [7] provide an excellent survey. Based on their literature review, they describe nine requirements to find interesting rules or associations: conciseness, generality/coverage, reliability, peculiarity, diversity, novelty, surprisingness, utility, actionability/applicability. These nine can be categorized into three groups: objective, subjective, and semantic measures.

Objective measures are based on data and do not take the user or application knowledge into account. An example of such a measure in biomedicine is an association that is evaluated for interestingness based only on the dataset under consideration. The probability of concepts occurring together is compared to probabilities for the entire dataset or adjusted based on appearance in other sources. For example, Jenssen [12] used MeSH terms assigned to abstracts to calculate the co-occurrence values of terms. Based on these values, they showed a network of genes extracted from Medline. Jayadevaprakash et al. [11] compared co-occurrence metrics for terms that were either based on MeSH terms assigned to an abstract or terms appearing in the abstract. They found the outcome to be equal. Narayanasamy et al. [19] focused on transitive associations of genes, proteins, or drugs found in Medline using co-occurrence measures. Later, this work was used in BioMap [17], which shows a graph of statistically defined (not NLP) directional relations.

Subjective measures add knowledge of the users' domain and/or background to the raw data. In biomedicine, such subjective measures are added by incorporating generalized background knowledge as encoded in resources such as the Gene Ontology (GO) (www.geneontology.org), the Unified Medical Language Systems (UMLS) (www.nlm.nih.gov/research/umls/), or even WordNet (http://wordnet.princeton.edu/). For example, Basu et al. [2] incorporated WordNet in their approach and found that their approach correlated as well with human judgments as the human judgments correlated with each other.

Less common are semantic measures, which include semantics and explanations of the patterns found. Hristovski et al. [10] exploit



Authorized licensed use limited to: to the Claremont Colleges!. Downloaded on January 22, 2010 at 12:25 from IEEE Xplore. Restrictions apply.

semantic relations extracted with SemRep and BioMedLEE to support a discovery system; they explicitly claim that semantic relations can support an explanation of discoveries. Fiszman et al. [5] use automatic semantic abstraction summarization to produce a graphical condensate from a set of MEDLINE abstracts. The method relies first on natural language processing (SemRep) that produces semantic predications. It then retains the most relevant predications (subject-relation-object) on a user specified topic based on principles of relevance, connectivity, novelty, and saliency [9].

## 3. Research Goal

In biomedical text mining, two main components are studied: Algorithms to find *good terms* and algorithms to find *good relations*. We focus on the second component: relations. Our overall goal is to show how NLP techniques can affect objective measures when mining relations between terms. Many measures and algorithms have been developed to select the most interesting relations from large sets, but we focus here on two common and fairly straightforward co-occurrence-based metrics: support and confidence. We chose these because they are easy to understand and allow us to show in this simple framework the impact of directionality of associations. Because we limit ourselves to relations between two terms (and not multiple terms, as in association rule mining), these metrics are very closely related to co-occurrence based visualizations of biomedical relations.

To evaluate the impact of directionality, several choices needed to be made with regard to the external resources (ontologies and thesauri), the parser (one used here), and the test bed. This work is the first of its kind and is a proof of concept. Therefore, we chose resources that were readily and publicly available. Future work will include additional ontologies, additional, optimized parsers and more focused collections that are tuned for biomedical researchers to use.

## 4. Methodology

### 4.1 Test Bed Creation

**Base Set Generation**. We downloaded a set of 193,384 abstracts from PubMed (www.ncbi.nlm.nih.gov/entrez/), by searching with the keyword "depression" (November 2006). PubMed was chosen as a source since it contains most of the biomedical literature and because our processor (SemRep) readily accepts this type of input. We chose depression as our topic because it has an increasing number of publications each year (see Figure 1) and because it contains several subtopics (e.g., genetics research for depression, depression and exercise, drugs and depression) and so provides a broader domain to evaluate our approach than would a specific gene.

The set of citations (titles and abstracts) was submitted as a batch job online to SemRep [20] (http://skr.nlm.nih.gov), which identifies semantic propositions in biomedical text. SemRep propositions are of the form subject-predicate-object and are based on the UMLS. Arguments (subject and object) are Metathesaurus® concepts, and predicates are permissible UMLS Semantic Network relations between concept types. During processing, SemRep relies on the SPECIALIST lexicon [18] and a biomedical part-of-speech tagger [22]. MetaMap [1], a biomedical named entity recognizer, then matches terms to concepts from the UMLS Metathesaurus® and determines the semantic type for each concept. Concepts are identified as arguments in a proposition using syntactic dependencies and semantic constraints imposed by the UMLS Semantic Network

The full-fielded output of SemRep provides predications, entities, and the original text together with its PubMed identification and mappings to the UMLS. A predication can be thought of as a normalized pattern found in the literature and consists of a triplet. For example, the sentence "*27 children with neuroblastoma were treated with 131I-Metaiodobenzylguanidine (MBIG)*" produces the predication "*3-Iodobenzylguanidine TREATS Central neuroblastoma*."

**Table 1: initial set of predications and final test bed**

| | |
|---|---|
| SemRep Output | |
| Total Predications : | 1,236,390 |
| Unique Subjects : | 34,475 |
| Unique Objects : | 29,930 |
| Unique Relations : | 83 |
| Test Bed | |
| Total Predications : | 126,404 |
| Unique Subjects : | 6,364, |
| Unique Objects : | 3,246 |
| Unique Relations : | 6 |



Authorized licensed use limited to: to the Claremont Colleges!. Downloaded on January 22, 2010 at 12:25 from IEEE Xplore. Restrictions apply.

**Table 2: example predications extracted by SemRep with the original text**

| Predication Type | | Predication Example | | | Original Sentence |
|---|---|---|---|---|---|
| Predicate | Freq. | Subject Concept | Predicate | Object Concept | |
| Examples of Correct Predications | | | | | |
| USES | High | Therapeutic Procedure | USES | Fluoxetine | Fluoxetine treatment of obsessive-compulsive disorder. |
| PREVENTS | Low | Adenosine | PREVENTS | Cell Injury | Adenosine protects against cellular damage and dysfunction under several adverse conditions, including inflammation. |
| DISRUPTS | Low | (1-6)-alpha-glucomannan | DISRUPTS | Mycotoxicosis | Efficacy of esterified glucomannan to counteract mycotoxicosis in naturally contaminated feed on performance and serum biochemical and hematological parameters in broilers. |
| Examples of Incorrect Predications | | | | | |
| PROCESS OF | High | Depressive Disorder | PROCESS OF | Population Group | Present at a frequency of 7.2%, the IL-2-Rbeta G245R was identified in a population of Eastern Sudan exposed to a severe outbreak of visceral leishmaniasis (VL), a disease associated with a marked depression of T-cell antigen-specific responses. |
| USES | Low | 4-Butyrolactone | USES | Ethanol | The enhancement of the dopa formation in dopaminergic neurons induced by GBL was markedly attenuated after chronic ethanol treatment. |

We separated the predications, entities, and original text into three separate files that can be loaded into a SQL Server database (The Perl script is available online). This script also removes sentences with irrelevant information, e.g., *This article contains Supplementary Material available at*, and the predications based on such sentences. We maintain a list with such end-of-abstract sentences that are unwanted. The top section of Table 1 provides an overview of the resulting set.

**Accuracy Estimation**: Because SemRep is under development, some predications are extracted with higher accuracy than others. Precision has earlier been estimated at 83% for the ISA predication [20] or ranging from 53% to 92% for other subsets of predications [5, 21]. Our goal was to use a highly accurate set of predications.

To construct this smaller and accurate test bed, we first evaluated different SemRep predications. From our set of abstracts, SemRep extracted 83 different predication types, e.g., ISA, TREATS, etc. Each could be found with and without negation, resulting in 166 combinations. For each such combination, we selected up to 5 examples that occurred very frequently across different abstracts and up to 5 examples that occurred infrequently across abstracts (some predication types had fewer than 5 examples). This provided us with 1,177 predications which were evaluated by the authors. Each predication was evaluated by one author.

When there was doubt about its correctness, the authors discussed and came to an agreement. A predication was considered incorrect when there was an incomplete noun phrase, incorrect match to a UMLS concept or incorrect relationship. Table 2 shows a few examples of correct and incorrect predications.

**Subset Creation.** Table 1 provides an overview of the size of our test set. To select the final subset, we employed four criteria. These limitations were chosen so that we would have a concise, manageable subset of predications to work with:

1) We excluded predications in which the subject was identical to the object or in which the subject or object was not clearly a noun (e.g., when it was a number).
2) We selected only predications whose subject or object had a UMLS semantic type relevant to genomics. This was done to avoid general



Authorized licensed use limited to: to the Claremont Colleges!. Downloaded on January 22, 2010 at 12:25 from IEEE Xplore. Restrictions apply.

relationships, such as "doctor – treats – patient," which would be filtered out by text mining algorithms. The semantic types that were acceptable for the subject and object were: Amino Acid, Peptide, or Protein; Acid Sequence; Biologically Active Substance; Chemical; Chemical Viewed Functionally; Chemical Viewed Structurally; Cell or Molecular Dysfunction; Element, Ion, or Isotope; Gene or Genome; Hormone; Immunologic Factor; Inorganic Chemical; Lipid; Nucleic Acid, Nucleoside, or Nucleotide; Neuroreactive Substance or Biogenic Amine; Nucleotide Sequence; Organic Chemical; Pharmacologic Substance; Receptor; Steroid.

3) We selected predications from categories which had 100% accuracy in our sample and which also demonstrated clear directionality. For example, the predication "X stimulates Y" shows clear directionality from x to y, whereas this is not as clear with "X interacts with Y." This resulted in 6 acceptable relations: "administered to," "isa," "treats," "stimulates," "part of," and "uses."
4) We excluded negated predication for this study to keep our test bed small and uniform.

### 4.2 Graph Manipulation

In most approaches, interestingness measures are first calculated for term-term combinations and then a subset of such relations is selected and displayed. In contrast to this and to demonstrate the impact of directionality, we first select the graphs and then evaluate them for interestingness. We built each sample graph around a central concept. This central concept represents the entry point that a researcher would look at in a large, comprehensive graph, i.e., the starting point for that researcher to explore the graph. We do not limit the number of nodes that can become part of the graph. We define a "graph" as consisting of the following:

- A central concept, which is randomly chosen from all our concepts.
- A first ring (inner ring) of associated concepts. These are concepts that are directly related to the central concept in the test bed. This forms a first set of predications with the central concept as either subject or object.
- A second ring (outer ring) of associated concepts, which are associated with a concept from the first ring. This forms a second set of predications with a subject or object that belongs to the inner ring.

We distinguish four types of graphs. Figure 3 shows an overview of all four. For each graph, there is a central concept, inner ring, and outer ring of concepts. Each type of graph differs in the required directionality of the associations between concepts. The first graph (A) is the baseline and does not limit the predications based on directionality. The second and third are similar, with predications that 'point' in the same direction, either away from the central concept (B) or towards the central concept (C). The last type of graph (D) contains predications that point from the central concept to another concept and from the outer ring to the inner ring concepts.

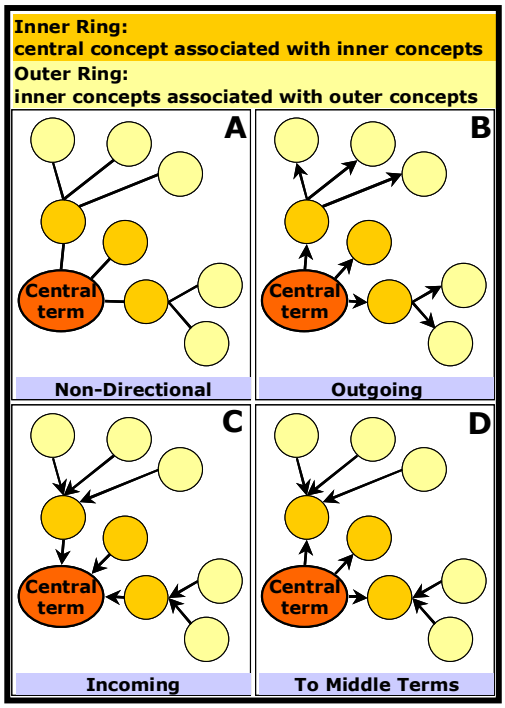


**Figure 3: A) baseline graph without directionality, B) outgoing predications, C) incoming predications and D) predications towards the inner ring**

### 4.3 Measures: Support and Confidence

Our goal is to show the impact of directionality. Toward this end, we selected the simplest interestingness measures: support and confidence. Support for a predication represents how often a relation between concepts can be found in an entire set: commonness of a relation between terms in the entire collection.



Authorized licensed use limited to: to the Claremont Colleges!. Downloaded on January 22, 2010 at 12:25 from IEEE Xplore. Restrictions apply.

Confidence in a predication shows the certainty that the second element follows when the first is present: correctness of a relation between terms in the entire collection. To take directionality into account, we adjusted support and confidence (we ignore the predicate or verb in the formulas):

*Non-Directional (Graph A):*
Support X – Y : n (XY) / N
Confidence X – Y: n (XY) / n (X)

n(XY): nr. of predications with X and Y
n(X): nr. of predications with X
N: total nr. of predications

*Directional (Graph B, C, D):*
Support X → Y : $n(X_1Y_2) / N$
Confidence X → Y: $n(X_1Y_2) / n(X_1)$

$n(X_1Y_2)$: nr. of predications with X as subject and Y as object
$n(X_1)$: nr. of predications with X as subject
N: total nr. of predications

Support and confidence are calculated for each predication. The results provide an overview of the average support and confidence for each type of graph and for both the inner and outer ring.

## 5. Results

For each type of graph, we randomly selected 100 concepts from the collection to serve as central concepts around which to build graphs. We used 100 small graphs instead of one large graph to avoid bias due to selection of a particular central concept.

### 5.1 Descriptive Statistics.

Table 3 provides an overview of the central concept and its appearance in the test bed. For each type of graph, we calculated how often the central concept appeared in the predications as either a subject or object (central concept frequency). On average, these concepts appeared between 12 and 22 times in a relation, however the variance is large, with some concepts appearing several hundred times, for example, venlafaxine appears 637 times, and many other concepts appear only once, for example Carbamyl Phosphate. The differences between the means are not statistically different. Although the large standard deviation in comparison to the means may indicate a non-normal distribution, analyses of variance are fairly robust against this deviation and so we used them for evaluation.

**Table 3: descriptive statistics for the central concepts**

| | | Central Concept Frequency | | | |
|---|---|---|---|---|---|
| Graph Type | N | Avg. | Min. | Max. | St.Dev |
| Non-Direct.: | 100 | 12 | 1 | 334 | 39 |
| Outgoing: | 100 | 16 | 1 | 318 | 46 |
| Incoming: | 100 | 22 | 1 | 466 | 66 |
| To Middle: | 100 | 16 | 1 | 290 | 42 |
| Overall: | 400 | 16 | 1 | 466 | 49 |

Table 4 shows the descriptive statistics for the inner ring, or the nodes connected to the central concept. The differences in means are not significant. However, as expected, there is a large standard deviation, indicating that some nodes appear in many relations. Some central concepts in our sample had only a few connections, while others had more than one hundred connections.

**Table 4: descriptive statistics for the inner ring (central – inner concepts)**

| | | Number of elements | | | |
|---|---|---|---|---|---|
| Graph Type | N | Avg. | Min. | Max. | St.Dev |
| Non-Direct.: | 100 | 5 | 0 | 91 | 11 |
| Outgoing: | 100 | 3 | 0 | 66 | 9 |
| Incoming: | 100 | 5 | 0 | 139 | 17 |
| To Middle: | 100 | 5 | 0 | 146 | 16 |
| Overall: | 400 | 4 | 0 | 146 | 14 |

Table 5 shows the descriptive statistics of the next ring of relations: the number of nodes on the outer ring or the number of connections from the inner to the outer ring. In this case, there was a main effect for graph type, $F(3,1) = 26.631$, $p < .001$. Differences between non-directional and outgoing and incoming graphs were significant at $p < .001$ level and at $p < .05$ level for the non-direction – to middle comparison (Bonferroni adjusted). There was no difference in means between outgoing and incoming graphs.

**Table 5: descriptive statistics for the outer ring (inner – outer concepts)**

| | | Number of elements | | | |
|---|---|---|---|---|---|
| Graph Type | N | Avg. | Min. | Max. | St.Dev |
| Non-Direct.: | 100 | 884 | 0 | 8,631 | 1,549 |
| Outgoing: | 100 | 2 | 0 | 157 | 16 |
| Incoming: | 100 | 2 | 0 | 47 | 7 |
| To Middle: | 100 | 1,499 | 0 | 15,996 | 2,373 |
| Overall: | 400 | 597 | 0 | 15,996 | 1,548 |



Authorized licensed use limited to: to the Claremont Colleges!. Downloaded on January 22, 2010 at 12:25 from IEEE Xplore. Restrictions apply.

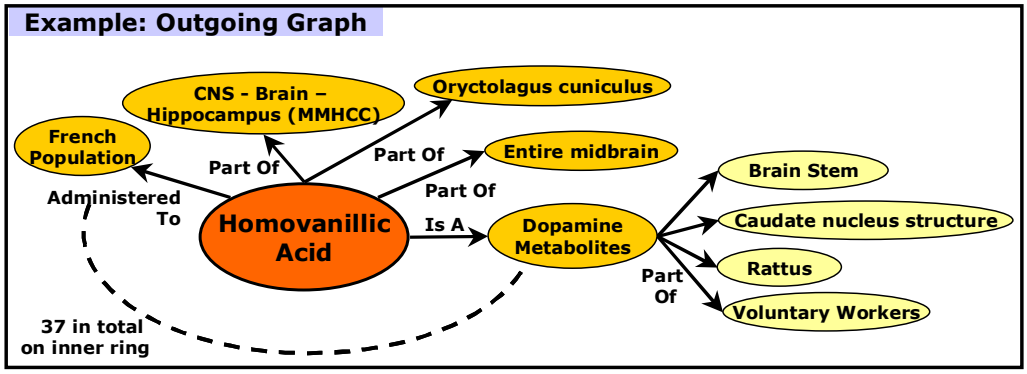


**Figure 4: partial example graph for outgoing predications**

Figure 4 shows an example outgoing graph for the central concept "Homovanillic Acid". There are 37 concepts associated with it that appear on the inner ring: 2 have the relation "administered to" (one shown), 1 has the "is a" relation (shown) and the 34 others have the "part of" relation (3 shown). In total, four of the inner ring concepts had outer ring concepts associated with them (shown only for Dopamine Metabolites).

Additional parsers will increase the number of different relations in the graphs. Ontological knowledge used for pre-selection or made available in the user interface will allow researchers to tune the graph, e.g., by focusing on general or specific concepts such as "Human" versus "French Populuation". Additional visualization can then also be added for different node types and frequency among others.

## 5.2 Support and Confidence.

Figure 5 shows average confidence for the different graphs for the two rings. In general, confidence is much higher for the inner ring than for the outer ring. For both the inner and outer ring, directionality plays a significant role.

We found a significant main effect for graph type for the inner ring predications, $F(3,1) = 9.460$, $p < .001$ with the incoming graph type significantly different from all three others at $p < .001$ (Bonferroni adjusted).

We also found a significant main effect for graph type for the outer ring, $F(3,1) = 17.917$, $p < .001$. In this case, post hoc comparisons showed that only the difference between non-directional and incoming graph types are significant ($p < .001$).

Figure 6 shows average support for the different types of graphs. As with confidence, directionality plays a significant role in support. We found a significant main effect of graph type on support for the inner ring predications, $F(3,1) = 5.589$, $p < .01$. Post hoc comparisons showed that the differences between non-directional and the other three graph types were significant, $p < .05$ (Bonferroni adjusted).

The second main effect for the outer ring predications was also significant, $F(3,1) = 59.489$, $p < .001$. Post hoc comparisons confirmed that the difference between incoming graphs and the three other graphs was significant at $p < .001$ (Bonferroni adjusted).

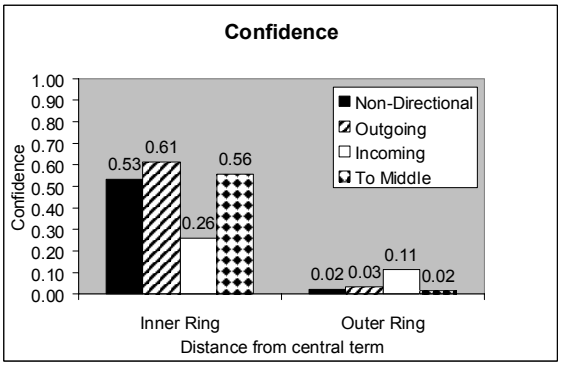


**Figure 5: average confidence for inner and outer ring for each graph type**

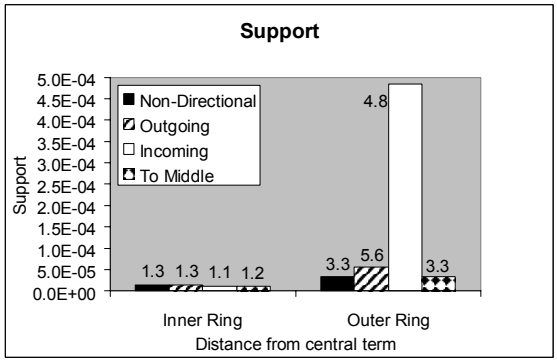


**Figure 6: average support for inner and outer ring for each graph type**



Authorized licensed use limited to: to the Claremont Colleges!. Downloaded on January 22, 2010 at 12:25 from IEEE Xplore. Restrictions apply.

## 6. Discussion

Since we randomly selected the starting concepts, the frequency of occurrence of these entities was not significantly different for the four graphs with 100 samples each. This shows that the differences found in each graph type cannot be attributed to the particular concepts in the graph but, as predicted, to the type of graph.

Although the number of nodes on the inner ring also did not differ statistically, the number on the outer ring did. This shows that further away from the central concept, the influence of the different type of graph becomes more pronounced.

Looking at all results, we see that non-directional relations show, as expected, lower support. The graph types in which predications of the inner and outer ring point in the same direction (outgoing and incoming) have the fewest number of nodes on the outer ring. However, the other numbers are not as easily described. More interesting are the differences in support and confidence. For both the inner and outer ring, we found that graph type has a large effect on confidence. Support was also significantly affected by graph type.

## 7. Conclusion and Future Work

This work used a test bed of relations between terms extracted from biomedical text. We developed the test bed so that all these relations showed clear directionality. We evaluated how this directionality affected the interestingness of a subset of relations. To measure this effect, we selected 100 sample graphs for four graph types: non-directional, incoming, outgoing, and to-middle. We then calculated support and confidence for each relation in each graph. These metrics were chosen because they are simple and well-known and are often used as building blocks in more intricate measures. We found that by manipulating directionality, we could manipulate interestingness.

This research needs to be followed up by two types of future work. On the one hand, it is necessary to look at the effect of directionality on larger graphs. And on the other hand, we need to establish what makes a graph interesting to researchers. Mixing and matching of different types of relations may lead to completely different graphs. Graphs with high confidence or support (or both) may provide a background of well-known information for researchers in which other graphs with low support but preferably high confidence may stand out and maybe lead to new discoveries.

## 8. Acknowledgments

M. Fiszman and T.C. Rindflesh were supported by the Intramural Research Programs of the National Institutes of Health, National Library of Medicine.

## 9. References

[1] A. R. Aronson, "Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program," presented at AMIA Symposium, 2001.

[2] S. Basu, R. J. Mooney, K. V. Pasupuleti, and J. Ghosh, "Evaluating the novelty of text-mined rules using lexical knowledge," presented at seventh ACM SIGKDD international conference on Knowledge discovery and data mining, San Francisco, California, 2001.

[3] C. Blaschke, M. A. Andrade, C. Ouzounis, and A. Valencia, "Automatic extraction of biological information from scientific text: protein-protein interactions," presented at 7th International Conference on Intelligent Systems for Molecular Biology, 1999.

[4] W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of text mining," *Communications of the ACM*, vol. 49, pp. 77-82, 2006.

[5] M. Fiszman, T. C. Rindflesch, and H. Kilicoglu, "Abstraction Summarization for Managing the Biomedical Research Literature," presented at Workshop on Computational Lexical Semantics (HLT-NAACL), 2004.

[6] S. S. Fuller, D. Revere, P. F. Bugni, and G. M. Martin, "A knowledgebase system to enhance scientific discovery: Telemakus," *Biomedical Digital Libraries,*, vol. 1, pp. 2, 2004.

[7] L. Geng and H. J. Hamilton, "Interestingness Measures for Data Mining: A Survey," *ACM Computing Surveys*, vol. 38, 2006.

[8] M. D. Gordon and R. 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*, vol. 47, pp. 116-128, 1996.

[9] U. Hahn and U. Reimer, "Knowledge-based text summarization: Salience and generalization operators for knowledge base abstraction," in *Advances in Automatic Text Summarization*, Mani and Maybury, Eds. Cambridge, London: MIT Press, 1999, pp. 215-232.

[10] D. Hristovski, C. Friedman, and T. C. Rindflesch, "Exploiting semantic relations for literature-



Authorized licensed use limited to: to the Claremont Colleges!. Downloaded on January 22, 2010 at 12:25 from IEEE Xplore. Restrictions apply.

based discovery,," presented at AMIA Annual Symposium, 2006.

[11] N. Jayadevaprakash, S. Mukhopadhyay, and M. Palakal, "Generating Association Graphs of Non-Cooccurring Text Objects using Transitive Methods," presented at 2005 ACM Symposium on Applied Computing, Santa Fe, New Mexico, USA, 2005.

[12] T.-K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig, "A literature network of human genes for high-throughput analysis of gene expression," *Nature Genetics*, vol. 28, pp. 21-28, 2001.

[13] A. Koike, Y. Niwa, and T. Takagi, "Automatic extraction of gene/protein biological functions from biomedical text," *Bioinformatics*, vol. 21, pp. 1227-1236, 2005.

[14] G. Leroy and H. Chen, "Genescene: An Ontology-enhanced Integration of Linguistic and Co-occurrence based Relations in Biomedical Texts," *Journal of the American Society for Information Science and Technology (Special Issue)*, vol. 56, pp. 457-468, 2005.

[15] G. Leroy, J. D. Martinez, and H. Chen, "A Shallow Parser Based on Closed-class Words to Capture Relations in Biomedical Text," *Journal of Biomedical Informatics*, vol. 36, pp. 145-158, 2003.

[16] Y. A. Lussier, T. Bolawski, D. Rappaport, Y. Liu, and C. Friedman, "PhenoGO: assigning phenotypic context to Gene Ontology annotations with natural language processing," presented at Pacific Symposium on Biocomputing, 2006.

[17] P. M, B. J, S. T, and H. S, "A comparative study of cells in inflammation, EAE and MS using biomedical literature data mining.," *Journal of Biomedical Science*, vol. 14, pp. 67-85, 2007.

[18] A. T. McCray, S. Srinivasan, and A. C. Browne, "Lexical methods for managing variation in biomedical terminologies," *Proc Annu Symp Comput Appl Med Care*, pp. 235-239, 1994.

[19] V. Narayanasamy, S. Mukhopadhyay, M. Palakal, and D. A. Potter, "TransMiner: Mining transitive associations among biological objects from text," *Journal of Biomedical Science*, vol. 11, pp. 864-873, 2004.

[20] T. C. Rindflesch and M. Fiszman, "The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text," *Journal of Biomedical Informatics*, vol. 36, pp. 462-477, 2003.

[21] T. C. Rindflesch, B. Libbus, D. Hristovski, A. R. Aronson, and H. Kilicoglu, "Semantic Relations Asserting the Etiology of Genetic Diseases," presented at AMIA Annual Symposium, 2003.

[22] L. Smith, T. C. Rindflesch, and W. J. Wilbur, "MedPost: a part-of-speech tagger for biomedical text,," *Bioinformatics*, vol. 20, pp. 2320-2321, 2004.

[23] P. Srinivasan and B. Libbus, "Mining MEDLINE for implicit links between dietary substances and diseases," *Bioinformatics*, vol. 20 (suppl. 1), pp. I290-I296, 2004.

[24] D. R. Swanson, "Raynaud's syndrome, and undiscovered public knowledge," *Perspectives in Biology and Medicine*, vol. 30,, pp. 7-18, 1986.

[25] J. M. Temkin and M. R. Gilder, "Extraction of protein interaction information from unstructured text using a context-free grammar," *Bioinformatics*, vol. 19, pp. 2046-2053, 2003.

[26] M. Weeber, H. Klein, A. R. Aronson, J. G. Mork, L. J.-V. d. Berg, and R. Vos, "Text-based discovery in biomedicine: the architecture of the DAD-system," presented at AMIA Annual Symposium, 2000.

[27] Y. T. Yen, B. Chen, H. W. Chiu, Y. C. Lee, Y. C. Li, and C. Y. Hsu, "Developing an NLP and IR-based algorithm for analyzing gene-disease relationships," *Methods of Information in Medicine*, vol. 45, pp. 321-329, 2006.



Authorized licensed use limited to: to the Claremont Colleges!. Downloaded on January 22, 2010 at 12:25 from IEEE Xplore. Restrictions apply.