Seeking Informativeness in Literature Based Discovery

Judita Preiss
University of Sheffield, Department of Computer Science
Regent Court, 211 Portobello
Sheffield S1 4DP, United Kingdom
j.preiss@sheffield.ac.uk

Abstract

The continuously increasing number of publications within the biomedical domain has fuelled the creation of literature based discovery (LBD) systems which identify unconnected pieces of knowledge appearing in separate literatures which can be combined to make new discoveries. Without filtering, the amount of hidden knowledge found is vast due to noise, making it impractical for a researcher to examine, or clinically evaluate, the potential discoveries. We present a number of filtering techniques, including two which exploit the LBD system itself rather than being based on a statistical or manual examination of document collections, and we demonstrate usefulness via replication of known discoveries.

1 Introduction and background

The number of publications in the biomedical domain has been observed to increase at a great rate, making it impossible for one person to read all, and thus potentially leaving knowledge hidden: for example, Swanson (1986) found one publication mentioning a connection between *Raynaud's Disease* and *blood viscosity* while another pointed out the effect of *fish oil* on *blood viscosity*, but there was no publication making the connection between *fish oil* and *Raynaud's Disease*. Automated approaches to knowledge discovery often set up the problem as outlined by Swanson; A being the source term (in this case *Raynaud's Disease*), with a possible target term, C, being specified (*fish oil*) and any connections between them form the linking, B, terms. If C is not specified, all possible hidden links from A are explored and discovery is classified as open. If both A and C terms are supplied, the discovery is closed and only any linking, B, terms are being sought.

Independent of how a connection between an A term and a B is defined (whether this is based on A and B co-occurring in the same title, in the same sentence or the same document, or some other relation), an obvious difficulty is the amount of data generated by a technique along these lines: with no filtering, a great number of connections will be made through terms such as *clinical study* or *patient*, and, if not also linked through other terms, these should be discarded. A number of approaches to term reduction have been explored.

Swanson and Smalheiser (1999)'s knowledge discovery system, Arrowsmith,[1] contains an increasing, currently 9,500 term (Swanson et al., 2006), stoplist, created semi-automatically.[2] Such a stoplist is unlikely to be complete – the list has grown from 5,000 (Swanson and Smalheiser, 1997) to 9,500 words (Swanson et al., 2006) and is likely to keep increasing. Over fitting is potentially an issue, in this case the list generated has been criticized for being tuned for the original *Raynaud–fish oil* discovery (Weeber et al., 2001). A word based stoplist also does not take into account the potential ambiguity of terms: one sense may be highly frequent and uninformative, guaranteeing it an appearance in the stoplist, while another sense may be rare but highly informative.

Instead of using words directly, it is possible to employ a (much smaller) controlled vocabulary: Medical Subject Headings (MeSH), consisting of 22,500 codes, are (mostly) manually assigned to each document indexed in Medline – even though multiple MeSH codes for a document are allowed, restricting to this set greatly reduces dimensionality. For example, Srinivasan (2004) uses MeSH based topic profiles to connect A to topics C via the most likely MeSH terms.

[1] Available at http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html

[2] Note that only 365 words of this stoplist are publicly available.

Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP 2014), pages 112–117,
Baltimore, Maryland USA, June 26-27 2014.

Keeping entire vocabularies is possible if topics are limited, for example, Fleuren et al (2011) extract statistics regarding gene co-occurrence, and restricts their hidden knowledge generation to biological mechanisms related to them.

Another difficulty in using word vocabularies is the necessary identification of multiwords, Weeber at al. (2001) avoid previously tried n-gram techniques (e.g. (Gordon and Lindsay, 1996)) by switching knowledge discovery to UMLS Concept Unique Identifiers (CUIs). Using MetaMap (Rindflesch and Aronson, 1994) to assign CUIs to texts discards non content words (CUIs only exist for concepts), resolves ambiguity and deals with multiwords in one, thus reducing the number of terms considered in later stages. Weeber et al. also exploit the broad subject categories that UMLS assigns to each CUI, which allow the authors to perform domain specific filtering to reduce dimensionality. This they do on a per search basis, tuning the filtering to the replication experiments presented.

Dimensionality reduction can also be performed at the relation level. Swanson's (1997) original work deemed two terms connected if they both appeared in the title of an abstract – titles were thought to be the most informative, and descriptive, part of each article. As the number of abstracts explored during the knowledge discovery process increased, and connections were extended to whole abstracts (rather than titles only), the amount of hidden knowledge generated increased dramatically and with it did the need for term and connection filtering.

Hristovski et al (2006) argue for filtering within the relation definition – co-occurrence does not provide any basis for a relation between two terms, no underlying semantic reason, and thus, as well as leading to many spurious links, it yields no justification for a hidden connection that is found. They extract subject-relation-object triples, with relations such as *treats* or *affects* forming their UMLS concept relations, leading to a much smaller number of (more accurate) relations to derive hidden knowledge from.

While re-ranking (placing the most 'useful' links at the top of the list) the resulting hidden knowledge is clearly valuable, removing terms from consideration prior to identifying hidden knowledge will reduce the computational load as well as avoid noisy hidden knowledge being produced and possibly accidentally being highly ranked.

We explore a number of filtering approaches including two novel techniques which can be integrated into any method designed using the Swanson framework, and we compare these against previously explored filtering methods. Section 2 outlines our knowledge discovery approach, Section 3 presents a number of filtering approaches with Section 4 discussing results based on replication of existing knowledge and Section 5 draws our conclusions.

2 Knowledge discovery system

There are two main components which define an LBD system created following the Swanson framework: the terms and the relations. Based on arguments presented in Section 1, our system employs UMLS CUIs as produced by SemRep (Rindflesch and Fiszman, 2003), a natural language processing system which identifies semantic relations in biomedical text.[3]

SemRep extracts relation triples from text by running a set of rules over the output of an underspecified parser. The rules, such as the mapping of *treatment* to TREATS, map syntactic indicators to predicates in the Semantic Network. Further restrictions are imposed regarding the permissibility of arguments, the viability of the given propositions, and other syntactic constraints, resulting in relations such as

- Epoprostenol TREATS Raynaud Phenomenon
- blood rheology DIAGNOSES Raynaud Disease

Each triple is also output with the corresponding CUIs.

All 29 non negative relations were extracted (such as AFFECTS, ASSOCIATED_WITH, INTERACTS_WITH, ...), while negative relations (such as NEG_AFFECTS, NEG_ASSOCIATED_WITH, NEG_INTERACTS_WITH, ...) were dropped. The extracted relations form the connections between CUIs: i.e., the set of linking CUIs B is created by following all SemRep links from the CUI A, which lead to C through another SemRep relation.

[3]In this work, the SemRep annotated Medline data, database semmedVER24 (processed up to November 2013) run over 23,319,737 citations to yield 68,000,470 predications, was downloaded from `http://skr3.nlm.nih.gov` and used throughout.

3 Filtering approaches

While employing CUIs (rather than words) eliminates non content words (thus immediately reducing noise), it does not eliminate CUIs corresponding to *patient*, *week*, *statement* ... We present, and in Section 4 evaluate (individually and in combination), four filtering approaches of which two are, to our knowledge, completely novel.

3.1 Synonyms

While not a filtering method under the usual definition, the identification of synonym CUIs and collapsing thereof results in the reduction of the number of CUIs being used (i.e. the technique filters out some CUIs).

A manual examination of the documents containing CUI C0034734, *Raynaud Disease*, revealed that some of the expected connections were missing and were linked to CUI C0034735, *Raynaud Phenomenon*, instead. The resulting hidden knowledge is greatly affected by the particular CUI chosen as the source term A, yet in this case, the two CUIs are synonymous. The MRREL related concepts file within UMLS contains pairs of CUIs within related relationships, including the SY (source asserted synonymy) relationship[4], and CUIs C0034734 and C0034735 appear in the SY relationship in this list. Identifying concepts within the SY relationship has the following advantages:

- Merging such synonyms into classes will allow the retrieval of more hidden knowledge if the multiple synonymous CUIs correspond to the start point, A (as in the case of *Raynaud Disease*).
- There will be potentially more hidden knowledge created if a multiclass CUI is a linking term (as A connected to C0034734 and C connected to C0034735 would not have been found to be connected if these were the only potential overlap).
- Synonymous hidden knowledge (and linking terms) will merge, reducing the amount of knowledge (and terms) to manually explore.

Merging synonyms into single CUI classes reduces the 561,155 CUIs present in UMLS to 540,440 CUI classes.[5]

[4]Due to the version of SemRep files used, UMLS 2013AA is employed throughout.

[5]Note that other MRREL related relationships were explored, but completing cycles lead to multiple extremely large equivalence classes.

3.2 Semantic types

The UMLS Semantic Network consists of 133 semantic types, a type of subject category, which is assigned to each CUI. Many of these categories are clearly unhelpful for knowledge discovery (for example, *geographic area* or *language*), and 70 semantic types are manually selected for removal (by examining the basic information about the relation, as well as the structure of the network and the CUIs assigned each semantic type). This removes a further 121,284 CUIs.

3.3 Discarding common linking terms

In some cases, a given CUI is clearly too general to be a useful linking term, but its semantic type contains more specific CUIs which should not be removed. Restricting semantic type filtering based on the depth within the hierarchy is also not a viable option, as UMLS is composed of different hierarchies, each with a different level of granularity and establishing an overall threshold would likely include general terms for some while discarding crucial terms for others. Therefore another approach is needed for these CUIs.

Along the lines of Swanson et al (2006), a stoplist can be built to contain such terms, without over-training for a particular discovery and without the need for manual intervention: we hypothesize that any CUIs which are linking terms more often than others can effectively form a stoplist.

The creation of this stoplist can be performed iteratively:

1. Start with an empty stoplist set S.
2. Create hidden knowledge based on SemRep connections between CUIs, removing any connections to CUIs in set S (the hidden knowledge is acquired from Medline articles published between 1865 and 2000).
3. Randomly select 10,000 hidden knowledge pairs, identify their linking CUIs, and add any linking CUIs appearing in more than $threshold$ of pairs to S (the value of *threshold* needs to be empirically determined).
4. If Step 3 increased the size of S, return to Step 2.

Note that since the training set is not designed for any particular discovery, this should not result in an over trained stoplist.

3.4 Breaking high frequency connections

The creation of a stoplist will always suffer from omissions and inclusions of CUIs that should not be filtered out in every instance. The last approach is based on a slightly different underlying idea: instead of finding frequently appearing terms, this approach bases its decisions on the number of terms a given term is connected to.

Two CUIs A and B are deemed connected if a (non negative) SemRep relation exists which links them. If A corresponds to a term such as *study* or *patient*, it is expected to be connected to a large number of CUIs. We hypothesize that terms which are so highly connected are likely to be relatively general terms, and so uninformative linking terms.

This gives rise to the following filtering options:

1. Break (discard) all connections to CUI A when the $C(A) >$ threshold.

2. Discard the connection between CUIs A and B when $min(C(A), C(B)) >$ threshold.

(Where $C(A)$ represents the number of CUIs linked to A, and the threshold needs to be empirically determined.)

Method 1 effectively forms a stoplist of highly connected CUIs, but method 2 is different: only connections satisfying the condition are broken while A remains under consideration. This allows filtering method 2 to leave a frequently connected term to be a linking term for a rare term (unlike method 1, which would discard such a term).

4 Results

Swanson's original discoveries (Swanson, 1986; Swanson, 1988) were verified through clinical trials and evaluation of LBD systems often involves replication of these discoveries (Gordon and Lindsay, 1996; Weeber et al., 2001). From literature, we identify seven separate discoveries to replicate (presented with the labels used in Table 1):

RD: Raynaud disease and fish oil (Swanson, 1986).

Arg: Somatomedin C and arginine (Swanson, 1990).

Mg: Migraine disorders and magnesium (Hu et al., 2006).

ND: Magnesium deficiency and neurologic disease (Smalheiser and Swanson, 1994).

INN: Alzheimer's and indomethacin (Smalheiser and Swanson, 1996a).

estrogen: Alzheimer's disease and estrogen (Smalheiser and Swanson, 1996b).

Ca^{2+}iPLA2: Schizophrenia and Calcium-Independent Phospholipase A2 (Smalheiser and Swanson, 1997).

The same subset of Medline as in each original discovery is employed for replication, and any abstracts containing a direct link between the two terms are removed (note that including these would not have affected the original discoveries as these only used titles) – thus any connections between A and C are necessarily hidden and require at least one linking term.

The Raynaud-fish oil and migraine-magnesium connections are the most commonly replicated discoveries, while the remaining discoveries are rarely explored. For CUI based investigations, this is likely due to the difficulty of selecting a representative CUI for the sought concepts. The second concept in the Schizophrenia and Calcium-Independent Phospholipase A2 connection is particularly tricky: UMLS suggests CUI C1418624 (PLA2G6 gene) as the most likely match, followed by CUI C2830173 (Calcium-Independent Phospholipase A2) as the second most likely. However, neither CUI is found in any relations in the given date range by SemRep. Closer examination reveals that the Ca^{2+}iPLA2 connections in the 1960-1997 Medline range are between CUI C0538273 (PLA2G6 protein, human). Not only does this highlight the difficulty of the replication task, it further motivates the need for a 'synonym' (or related concept) list.

The number of linking terms found between each pair of sought terms is presented in Table 1 (zero linking terms means the connection was not found) for a subset of the filtering results. ST represents semantic type filtering, HF the breaking of high frequency connections (a *min* subscript denoting the version which takes into account connectivity of both CUIs), together with the threshold value, and LT elimination of common linking terms, again with the relevant threshold value.

While the Raynaud-fish oil connection appears to be consistently produced by the system, Table 2 reveals the value of filtering: with no filtering, the two linking terms are pure noise and the connection should not be made. Employing UMLS syn-

	RD	Arg	Mg	ND	INN	estrogen	Ca^{2+}iPLA2
No filtering	2	235	78	98	370	500	7
Synonyms (Sy)	6	173	58	65	296	415	16
Sy & LT-200	3	145	48	56	265	0	16
Sy & HF-2900	6	149	56	52	243	0	14
Sy & HF_{min}-900	6	73	22	27	82	164	9
Sy & HF_{min}-400	6	25	5	8	25	65	8
Sy & ST	4	130	47	43	234	331	13
Sy & ST & LT-200	3	108	41	38	207	0	13
Sy & ST & HF-2500	4	120	47	38	205	0	13
Sy & ST & HF_{min}-900	6	73	22	27	82	164	9
Sy & ST & HF_{min}-400	4	28	6	12	30	73	6

Table 1: Number of hidden links found during replication

onyms adds genuine linking terms,[6] and restricting by semantic types drops the remaining general terms. Discarding common linking terms finds *antimicrobial susceptibility* to be a frequently used linking term, and it is also dropped. A great advantage of the technique can be seen when connections are made through hundreds of terms – in this case, higher thresholds (and thus more aggressive filtering) can be employed to reduce the number of linking terms to the most promising set. Should these not be sufficient, the threshold can be increased to produce more linking terms and as such, the burden on the user in checking a large number of linking terms when a hidden connection is suspected can be greatly reduced, without sacrificing connections should more be needed.

Term	NF	Sy	Sy ST	LT-200
acetylsalicyclic acid	×	✓	✓	✓
antimicrobial susceptibility	×	✓	✓	×
blood viscosity	×	✓	✓	✓
brain infarction	×	✓	✓	✓
patient	✓	✓	×	×
volunteer helper	✓	✓	×	×

Table 2: Linking term analysis for RD

For example, common linking term filtering removes the term *estrogen* from consideration as *therapeutic estrogen* is a commonly used linking term, making the *estrogen-AD* link impossible to find. Linking term frequencies (on a 10,000 pair sample) exceeding values from 50 to 200 (in increments of 50) were tested resulting in the removal of between 1,902 and 227 CUIs. *Therapeutic estrogen* appears in all the lists. Similarly, the CUI is dropped when high frequency connections are broken using the first technique, which is based on stoplists. This highlights the value of the second high frequency connection technique, which only discards particular connections (rather than CUIs) and *therapeutic estrogen* CUI remains a searchable CUI.

As shown, the system replicates most of the previously published discoveries with its main asset being noise reduction: the number of linking terms for a suspected connection (closed discovery) can be greatly reduced to remove spurious connections, with backoffs available to yield more connections should more be required. For novel applications (i.e. open discovery), the technique greatly reduces the amount of hidden knowledge generated from a source term A. For example, the amount of hidden knowledge generated from somatomedin C drops from 82,601 CUIs when no filtering is performed, to 3,005 CUIs with synonym, semantic type and breaking connections with frequency more than 200, which represents a great reduction for a user who is likely looking for a particular type of C term.

5 Conclusions and future work

We present and demonstrate the effectiveness of a number of filtering methods, including two novel techniques based on any LBD system built according to the Swanson framework – one approach based on stoplist methods, but requiring no manual intervention except for a user's selection of a threshold, and the second based on removing connections when these are deemed to be likely to

[6]Note that the merging of synonyms is achieved without the need to back off to general classes (e.g. (Srinivasan, 2004)), which have been observed to lead to connections based on "aboutness" rather than producing genuine hidden knowledge (Beresi et al., 2008).

contribute mainly noise. A great advantage of the second approach is shown to be the fact that terms are not directly discarded, as with a stoplist, and thus a fairly common term can remain a source term when required.

While the method is evaluated by replicating known discoveries, we suggest that the noise reduction performed is ultimately leading to a much more user friendly LBD system, and plan to investigate other evaluation approaches, such as timeslicing (Yetisgen-Yildiz and Pratt, 2009), as part of future work.

Acknowledgements

Judita Preiss was supported by the EPSRC grant EP/J008427/1: Language Processing for Literature Based Discovery in Medicine.

References

Ulises Cervino Beresi, Mark Baillie, and Ian Ruthven. 2008. Towards the evaluation of literature based discovery. In *Proceedings of the Workshop on Novel Evaluation Methodologies (at ECIR 2008)*, pages 5–13.

Wilco W. M. Fleuren, Stefan Verhoeven, Raoul Frijters, Bart Heupers, Jan Polman, René van Schaik, Jacob de Vlieg, and Wynand Alkema. 2011. Copub update: Copub 5.0 a text mining system to answer biological questions. *Nuclear Acids Research*, 39 (Web Server issue). doi:10.1093/nar/gkr310.

Michael D. Gordon and Robert K. Lindsay. 1996. Toward discovery support systems: a replication, reexamination, 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.

Thomas C. Rindflesch Hristovski D, Friedman C and Peterlin B. 2006. Exploiting semantic relations for literature-based discovery. In *Proceedings of the 2006 AMIA Annual Symposium*, pages 349–353.

Xiaohua Hu, Xiaodan Zhang, Illhoi Yoo, and Yanquing Zang. 2006. A semantic approach for mining hidden links from complementary and non-interactive biomedical literature. In *SDM*.

Thomas C. Rindflesch and Alan R. Aronson. 1994. Ambiguity resolution while mapping free text to the UMLS Metathesaurus. In J. G. Ozbolt, editor, *Proceedings of the Eigheeth Annual Symposium on Computer Appplications in Medical Care*, pages 240–244.

Thomas C. Rindflesch and Marcelo Fiszman. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6):462–477.

Neil R. Smalheiser and Don R. Swanson. 1994. Assessing a gap in the biomedical literature: Magnesium deficiency and neurologic disease. *Neuroscience Research Communications*, 15(1):1–9.

Neil R. Smalheiser and Don R. Swanson. 1996a. Indomethacin and Alzheimer's disease. *Neurology*, 46:583.

Neil R. Smalheiser and Don R. Swanson. 1996b. Linking estrogen to Alzheimer's disease. *Neurology*, 47:809–810.

Neil R. Smalheiser and Don R. Swanson. 1997. Calcium-independent phospholipase a2 and schizophrenia. *Arch Gen Psychiatry*, 55(8):752–753.

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

Don R. Swanson and Neil R. Smalheiser. 1997. An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence*, 91:183–203.

Don R. Swanson and Neil R. Smalheiser. 1999. Link analysis of MEDLINE titles as an aid to scientific discovery: Using Arrowsmith as an aid to scientific discovery. *Library Trends*, 48:48–59.

Don R. Swanson, Neil R. Smalheiser, and Vetle I. Torvik. 2006. Ranking indirect connnections in literature-based discovery: The role of medical subject headings. *Journal of the American Society for Information Science and Technology*, 57(11):1427–1439.

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

Don R. Swanson. 1988. Migraine and magnesium – 11 neglected connections. *Perpectives in Biology and Medicine*, 31(4):526–557.

Don R. Swanson. 1990. Somatomedin c and arginine: Implicit connections between mutually isolated literatures. *Perspectives in Biology and Medicine*, 33(2):157–186.

Marc Weeber, Rein Vos, Henny Klein, and Lolkje T. W. de Jong-van den Berg. 2001. Using concepts in literature-based discovery: Simulating Swanson's Reynaud – fish oil and migraine – magnesium discoveries. *Journal of the American Society for Information Science and Technology*, 52(7):548–557.

M. Yetisgen-Yildiz and W. Pratt. 2009. A new evaluation methodology for literature-based discovery. *Journal of Biomedical Informatics*, 42(4):633–643.