HiDE: A Tool for Unrestricted Literature Based Discovery

Judita Preiss
University of Salford
The School of Computing, Science & Engineering
Newton Building, Salford
Greater Manchester M5 4WT
J.Preiss@salford.ac.uk

Mark Stevenson
Department of Computer Science
University of Sheffield
211 Portobello
Sheffield S1 4DP
Mark.Stevenson@sheffield.ac.uk

Abstract

As the quantity of publications increases daily, researchers are forced to narrow their attention to their own specialism and are therefore less likely to make new connections with other areas. Literature based discovery (LBD) supports the identification of such connections. A number of LBD tools are available, however, they often suffer from limitations such as constraining possible searches or not producing results in real-time.

We introduce HiDE (Hidden Discovery Explorer), an online knowledge browsing tool which allows fast access to hidden knowledge generated from all abstracts in Medline. HiDE is fast enough to allow users to explore the full range of hidden connections generated by an LBD system. The tool employs a novel combination of two approaches to LBD: a graph-based approach which allows hidden knowledge to be generated on a large scale and an inference algorithm to identify the most promising (most likely to be non trivial) information.

Available at https://skye.shef.ac.uk/kdisc

1 Introduction

Literature based discovery (LBD) is an automatic technique addressing the ever increasing volume of research literature by inferring as yet unobserved connections. The approach was pioneered by Swanson (1986) who hypothesised a (hidden) connection between *Raynaud phenomenon* and *fish oil*, despite the fact that the two were not mentioned together in any publications. Swanson noticed that one publication linked *Raynaud phenomenon* to *blood viscosity* and another linked *blood viscosity* to *fish oil*, suggesting the trial of administering fish oil to Raynaud disease patients. LBD can be executed in one of two modes: closed or open discovery. In closed discovery, both A, the source term, and C, the target term, are specified, and only the linking terms (with relationships to both A and C) are sought, while open discovery explores a much larger space with only the source term being specified and all relationships being pursued (see Figure 1).

LBD has a range of applications including identification of potential treatments, drug repurposing and drug side effect prediction. However, in its general form LBD generates a vast number of hidden connections and the usefulness of existing open discovery systems, such as **Arrowsmith** (Swanson and Smalheiser, 1999), **Bitola**'s (Hristovski et al., 2006), **FACTA+** (Tsuruoka et al., 2008) or **Literome** (Poon et al., 2014), is often limited by heavy restrictions on the input, linking terms and output and/or time required to generate results.

2 Approach

HiDE combines two LBD approaches. To ensure a usable (rather than excessive) quantity of quality hidden knowledge, we combine: (1) the widely used *A-B-C model* introduced by Swanson (1986) which starts from a term, A, finds all terms B_i to which A is related, repeats the process to find all terms C_{ij} related to each B_i, and proposes any previously unconnected $A - C_{ij}$ as hidden knowledge, and (2) a

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

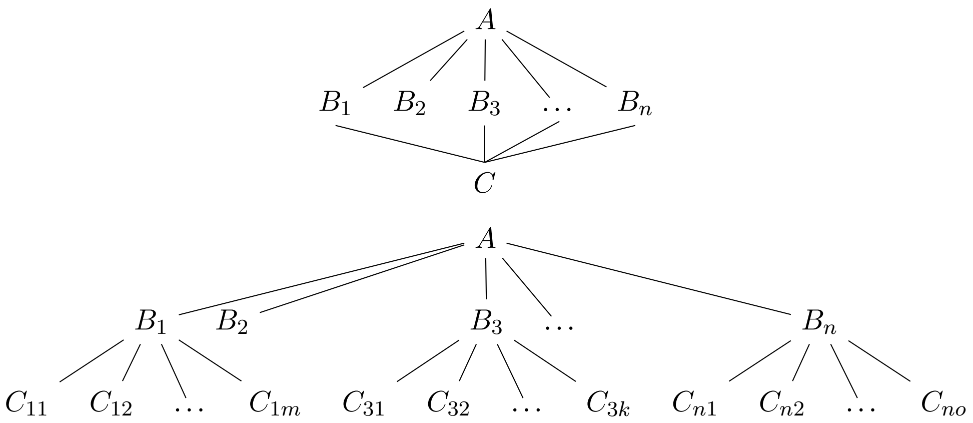

Figure 1: Closed discovery, both A and C specified (top), and open discovery, only A specified (bottom)

novel (to LBD) approach based on work in knowledge base completion which generates new connections by performing random walks through a knowledge base graph.

The *A-B-C model* is a useful approach for LBD but it can generate vast amounts of hidden knowledge potentially leading to the need for restrictions on the *B*/*C* terms and/or slow processing times. Exploiting techniques from graph theory (West, 2007), our LBD system (Preiss et al., 2015) uses the adjacency matrix M describing the graph formed from the connections between terms in a document collection: entry m_{ij} is a positive integer if a relation R is detected between terms t_i and t_j. If t_i and t_j are not related anywhere in the document collection, m_{ij} will be zero. Hidden knowledge in the document collection can then be identified by looking for non zero terms in the matrix generated by norm(M^2) − norm(M) where norm converts m_{ij} to 1 if $m_{ij} > 0$ and leaves it as 0 otherwise. This generates hidden knowledge connected via a single linking step and allows large amounts of hidden knowledge to be pre-computed.

The *graph model* is an inference system due to Lao et al (2011) based on the Path Ranking Algorithm, which performs random walks through a knowledge base graph. In our case, the knowledge base is constructed from the manually created triples (such as *X may treat Y*) listed in the Unified Medical Language System (UMLS) Metathesaurus. The system generates path up to length 2, and uses logistic regression to combine the paths to yield new connections.

Both LBD systems are applied to all PubMed abstracts published up to 30 April 2016: the linguistically motivated *subject-relation-object* triples (such as *X-treats-Y* or *X-affects-Y*) are extracted from a SemRep (Rindflesch and Fiszman, 2003) annotated 2016 version of PubMed (available as semmedVER26 download created using regular SemRep version 1.7 and UMLS 2016AA[1]) and used for the *A-B-C model*. UMLS 2016AA was used to obtain the manually created triples for the *graph model*. A range of filtering approaches are applied to reduce the volume of hidden knowledge (Preiss, 2014). Individually, the *A-B-C model* generated a total of 2,947,874,564 pairs of hidden knowledge, while the *graph model* yielded 198,295,133 pairs. The intersection of hidden knowledge pairs, 6,471,922 pairs, is presented within the interface, and the hidden knowledge pairs are ranked by the weights output by the *graph model*.

3 Online System

The approach described in Section 2 is implemented as a publicly available tool, HiDE (Hidden Discovery Explorer), which allows a user to interactively explore the hidden knowledge generated by an LBD system.

Interaction with HiDE begins with the user specifying a term of interest. HiDE then generates a list of potentially relevant UMLS CUIs from which the user selects one. The hidden knowledge available is grouped by UMLS Medical Subject Headings (MeSH) terms which provides types such as *disease*,

[1] `https://semrep.nlm.nih.gov/`

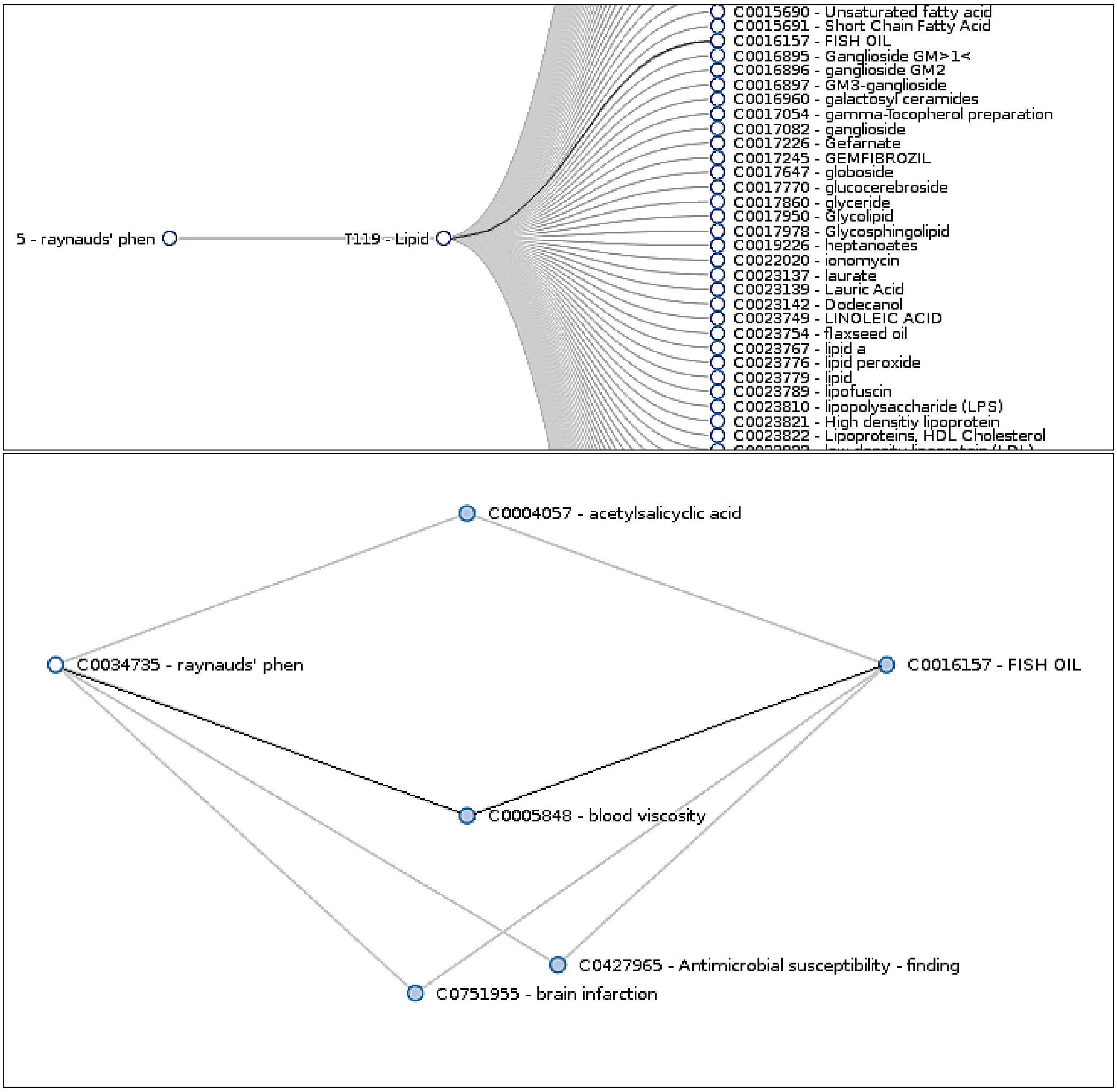
C0015690 - Unsaturated fatty acid
C0015691 - Short Chain Fatty Acid
C0016157 - FISH OIL
C0016895 - Ganglioside GM>1<
C0016896 - ganglioside GM2
C0016897 - GM3-ganglioside
C0016960 - galactosyl ceramides
C0017054 - gamma-Tocopherol preparation
C0017082 - ganglioside
C0017226 - Gefarnate
C0017245 - GEMFIBROZIL
C0017647 - globoside
C0017770 - glucocerebroside
C0017860 - glyceride
C0017950 - Glycolipid
C0017978 - Glycosphingolipid
C0019226 - heptanoates
C0022020 - ionomycin
C0023137 - laurate
C0023139 - Lauric Acid
C0023142 - Dodecanol
C0023749 - LINOLEIC ACID
C0023754 - flaxseed oil
C0023767 - lipid a
C0023776 - lipid peroxide
C0023779 - lipid
C0023789 - lipofuscin
C0023810 - lipopolysaccharide (LPS)
C0023821 - High densitiy lipoprotein
C0023822 - Lipoproteins, HDL Cholesterol
5 - raynauds' phen
T119 - Lipid
C0004057 - acetylsalicyclic acid
C0034735 - raynauds' phen
C0016157 - FISH OIL
C0005848 - blood viscosity
C0427965 - Antimicrobial susceptibility - finding
C0751955 - brain infarction

Figure 2: Raynaud phenomenon open discovery: **top** shows the first page of *lipid* hidden knowledge from *C0034735 – Raynaud Phenomenon* generated from publications between 1960 and 1985 highlighting Swanson's fish oil connection, **bottom** the linking terms between *C0034735 – Raynaud Phenomenon* and *C0016157 - fish oil* with the highly cited *blood viscosity* link highlighted

enzyme and *gene*. The user selects a MeSH term, which allows them to filter the result set to MeSH terms of relevant to them while also reducing the number of results returned, and the hidden knowledge generated from the original CUI is presented. Users can view hidden knowledge in increments of 100 pairs and linking terms in increments of 50.

3.1 Implementation Details

HiDE is a web-based system in which all rendering is achieved using the D3 JavaScript library. Hidden knowledge is generated offline and stored in a MySQL database which the interface accesses using PHP. Linking terms for a selected pair of CUIs are computed in real time. All results are cached to ensure subsequent access for the same knowledge pair will be virtually instant.

4 Example

Figure 2 presents the output of HiDE when replicating the connection between *Raynaud* and *fish oil* (Swanson, 1986) from 1960-8 Medline publications using the matrix method only (as the inference method would require a UMLS from 1968 which does not exist). The top portion of Figure 2 shows a zoomed in section of the hidden knowledge generated by HiDE by entering the search term *raynaud*, selecting the CUI *C0034735 – Raynaud Phenomenon* and then the MeSH term *lipid*. The figure shows that the link to the C term *fish oil* is found by HiDE (this link is highlighted). Selecting this CUI reveals the B term(s) via which the hidden knowledge was established; the bottom of Figure 2 shows the linking terms between *Raynaud* and *fish oil*, demonstrating that HiDE finds the frequently cited link via *blood viscosity* (highlighted).

5 Conclusion

We present HiDE, an LBD tool suitable for exploring hidden knowledge generated by an LBD system including linking terms. Rather than imposing a filtering by design, HiDE does not restrict the hidden knowledge presented to the user while allowing them to quickly drill down to MeSH terms of interest and thus carry out their own 'filtering'. Using a novel combination of two LBD approaches – a graph-based approach and an inference algorithm – the most promising information is computed off line, thereby enabling fast response times to queries and allowing users to fully explore the information generated.

Acknowledgements

The work described in this paper was funded by the Engineering and Physical Sciences Research Council (EP/J008427/1).

References

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

Ni Lao, Tom M. Mitchell, and William W. Cohen. 2011. Random walk inference and learning in a large scale knowledge base. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 529–539.

US National Library of Medicine. Semantic knowledge representation. https://semrep.nlm.nih.gov/. Accessed: 31-07-2017.

Hoifung Poon, Chris Quirk, Charlie DeZiel, and David Heckerman. 2014. Literome: PubMed-scale genomic knowledge base in the cloud. *Bioinformatics*, 30(19):2840–2842.

Judita Preiss, Mark Stevenson, and Robert Gaizauskas. 2015. Exploring relation types for literature-based discovery. *Journal of the American Medical Informatics Association*, 22:987–992.

Judita Preiss. 2014. Seeking informativeness in literature based discovery. In *Proceedings of BioNLP 2014*, pages 112–117.

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.

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

Y. Tsuruoka, J. Tsujii, and S. Ananiadou. 2008. Facta: a text search engine for finding associated biomedical concepts. *Bioinformatics*, 24(21):2559–2560.

Douglas B. West. 2007. *Introduction to Graph Theory*. Prentice Hall.