

联系强弱对生物医学文献知识发现的影响^{*}

闵波 张莉 杨帆 雷权 张云宏

(新疆军区总医院信息科 乌鲁木齐 830000)

〔摘要〕 分析间接关联的联系强弱在生物医学文献知识发现过程中的不同表现,明确关联性强弱对文献知识发现的影响与作用,表明通过强关联发现潜在关联的准确率显著高于由弱关联所获得的准确率,强关联性可作为提升文献知识发现应用效果的重要因素。

〔关键词〕 文献挖掘;知识发现;弱联系;强联系

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2017.05.002

The Effect of Strong - weak Connection on Medical Literature Knowledge Discovery MIN Bo, ZHANG Li, YANG Fan, LEI Quan, ZHANG Yun - hong, General Hospital of Xinjiang Military Region, Urumqi 830000, China

〔Abstract〕 The paper analyzes different expressions of indirectly associated strong - weak connection in the process of biomedical literature knowledge discovery, defines the effect and function of strong - weak connection on literature knowledge discovery, and indicates that the accuracy rate of the discovery of potential connection through strong connection is obviously higher than the accuracy rate obtained through weak connection, and strong connection can serve as an important factor for improving the application effect of literature knowledge discovery.

〔Keywords〕 Literature mining; Knowledge discovery; Weak connection; Strong connection

1 引言

生物医学研究积累了大量的文献数据,其中记载了大量不同类型的知识,为了向生物医学科研人员提供有序化的信息、挖掘文献中的潜在知识,研究文献知识发现的新方法、新技术、新系统,是生物医学信息学与知识服务的一个重要方向。传统检索平台能够辅助快速查找文献,但候选文献的数量庞大,要想从中获得有效的知识,需要人工阅读分

析,这是一项费时费力甚至是不可能实现的工作。美国芝加哥大学情报学教授 Don. R. Swanson 创立的非相关文献知识发现方法可以在大量医学文献中揭示学科间的隐藏联系,形成合理的假设来指导科学实践^[1]。基于文献的知识发现研究通过潜在的间接关联挖掘来获取新的科学假设^[2],即概念 A 与 C 之间虽然没有任何的文献报道的关系,但若 A 与 B 、 B 与 C 分别存在有效的关系,那么 A 与 C 之间可能通过共同的 B 存在某种联系,这可能是一个新的科学发现^[3]。如有文献报道雷诺病人(C)有典型的血液和血管相关的血黏度、血小板凝集度升高、血管收缩等生理改变 B ,同时又有文献报道鱼油 A 及其活性成分可以降低血黏度、血小板凝集度与引起

〔修回日期〕 2017-05-23

〔作者简介〕 闵波,技师;通讯作者:张莉,副主任医师。

血管舒张，由此得出鱼油 A 可能对雷诺病 C 有治疗作用的科学假设。

间接关联的联系评价是生物医学文献挖掘研究的热点问题，是文献知识发现系统实现的核心组成，有效地评价 A 、 B 、 C 3 者间的联系，才可以自动识别与形成新的科学假设。目前在生物医学社区已经存在大量文献知识发现系统，如 FACTA+^[4]、Literome^[5]、CoPub^[6] 等，其中涉及的各类间接联系评价的方法很多，但尚未有一个领域科学家公认的标准方法，多是为了适合相应的系统应用而提出^[7]。因此，从不同角度探索文献知识发现的影响因素与内在规律，对寻找新的评价方法具有非常重要的意义。本文主要通过分析强关联与弱关联在文献知识发现过程中间接关联评价时的不同表现，以探索间接关联的相关性强弱对知识发现的作用与影响。

2 研究方法

2.1 概述

生物医学文献知识发现过程主要包括概念实体 A 、 B 和 C 的识别、关联的抽取以及间接关联的相

关度计算。对于一个文献知识发现系统，给定某一概念 A ，通过计算 $A-B-C$ 之间的关联强度，获得按关联强度由大到小的有序列表^[8]。当返回的候选结果的数量较大时，排在前面的结果通常是最符合用户需求的，因此，给定每一个 $A-B-C$ 关系链，计算其关联性的一个分值 S_{AC} ，利用 S_{AC} 对结果集进行置优排序以筛选候选关联，实现对所有候选结果集从大到小排序之后，在有序的结果列表中，确保排序靠前的多是全局关联性评价较好的结果，即计算分值最高的 L 条间接关系链 $A-B_i-C_{j(1).....(L)}$ 的准确率 P 。由于不同研究内容的热度不同，不同概念所覆盖的文献差异很大，因此，为了评价知识发现的总体效果，通常随机给定一定数量的种子词集 A_{seed} ，计算所有种子词项获得的总体有序集的准确率 P ， P 越大效果越好。如图 1 所示，为了比较联系强弱对发现潜在关联的不同表现，将测试文献数据集按时间从小到大顺序分割为训练集与测试集，在 T_{train} 时期的训练集中随机选择 50 个词项 A_{random} 作为种子词项集；基于 T_{test} 时期的测试集，按照关联强度顺序（由小到大）与逆序（由大到小）排序，计算与比较两种情况下的准确率 P_{weak} 与 P_{strong} 。

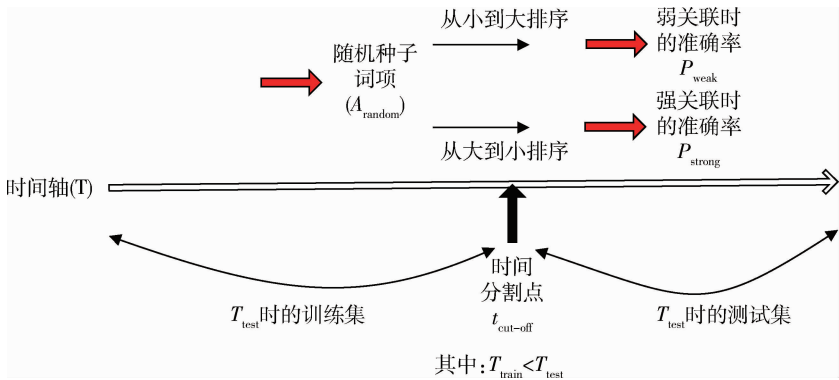


图 1 强弱关联在间接关联发现过程中不同表现的比较

2.2 关联建模

利用图对关联知识进行建模是目前相关领域最常用的方法。图是由节点集 V 和节点间的关系集合 E (边的集合) 组成的一种数据结构，可以用二元组定义为： $G = (V, E)$ ，其中 $V = \{v_1, v_2, \dots, v_n\}$ ， $E = \{e_1, e_2, \dots, e_m\}$ ，任意一条边对应一个

节点的二元组： $e_x = (v_i, v_j)$ 。将文献集中的内容相关性转化为基于关联信息存在的图结构数据模型，即根据文献中的概念实体及其关联信息，对文献中所蕴含的知识进行网络结构化，构建文献关联知识网络^[9]。在生物医学文献关联知识网络中，节点是各种从生物医学文献中提取而来的实体，如基因、蛋白质、化合物或疾病等，边是实体之间的关

联。关联知识网络把文献集中的知识以网络形式表示出来,既表示出知识之间的联系,也过滤了冗余知识。本文对生物医学文献关联知识网络中的实体关系通过共现法进行抽取。共现法实际应用最为广泛,认为当两个词语共现于一定的语境中时,词语之间可能存在一定的语义相关性,具体使用包括不同层次的分析单元,如单词、句子、摘要或者全文,其中以句子为最大分析单元是最常见的选择^[10]。共现能够描述实体对象之间的关联,但是这种关联对于不同类型的共现实体来说,所代表的生物医学含义有时会存在很大的差异性,如某两个实体可能只共现一次,也存在明确的语义关系;频繁共现的实体间虽具有很强的关系,却可能对知识发现的作用不大,而某些非频繁的异常关联对发现潜在联系可能更具实际意义。笔者利用 NLTK 自然语言处理包实现文献文本处理与关联提取,用于后续的数据测试,主要包括分词、词性标注、分块及名词短语提取,利用医学主题词表(Medical Subject Headings, MeSH)辅助进行词干化,对于可识别的实体而言,在文本数据中识别并定位实体名称后,如果两个概念实体 A 与 B 在同一句子中共现,无论它们共现的次数多少,记为关系 AB ,基本步骤如下:(1)基于自然语言的方法识别出句子的实体 NP 及其位置。(2)如果在同一个句子中得到的实体按其在句子中的顺序依次为 $NP1$ 、 $NP2$ 、 $NP3$,则得到关联: $(NP1, NP2)$, $(NP1, NP3)$, $(NP2, NP3)$ 。共现法获得的直接关联尽管是未明确的语义关系,会带来一定的噪声与增加总的关联数量,但通过置优排序后靠前的关联,相对来说更可能是存在潜在联系的间接关联,作为一项科学假设,更值得去做进一步的分析,从靠前区间富集的趋势来观察,不会影响结果的整体表现。

2.3 准确率 P 的计算

为了计算准确率 P ,将给定文献数据集按时间分成训练集和测试集,分别建立训练网络 $G_1 = (N_1, E_1)$ 和测试网络 $G_2 = (N_2, E_2)$ 。从 N_1 中随机选择 X 个词作为种子词项集 A ,其中 A 取训练网络与测试网络中共同拥有的词项,即 $A \in N_1 \cap N_2$ 。

(1) 在训练网络 G_1 中,以种子集 A 中的节点为起点提取其间接节点,得到间接节点集 C ,计算所有关系链 $(A - B_{\text{train}} - C)$ 的一个分值 S_{AC} ,对结果集 C 按 S_{AC} 值从大到小排序,取有序结果集 C_{Sorted} 中前 L 个词项,得到 $C_{\text{Sorted_TopL}} = \{c_1, c_2, \dots, c_L\}$ 。(2) 在测试网络 G_2 中,以种子节点集 A 中的节点为起点提取其直接关联节点,得到关联节点集 B_{test} 。(3) 计算有序结果集 C_{Sorted} 前 L 个词项集 $C_{\text{Sorted_TopL}}$ 的准确率 P :

$$P = \frac{|C_{\text{Sorted_TopL}} \cap B_{\text{test}}|}{|B_{\text{test}}|}$$

式中: $C_{\text{Sorted_TopL}} \cap B_{\text{test}}$ 指 $C_{\text{Sorted_TopL}}$ 和 B_{test} 的交集,即共同拥有的词项, $|C_{\text{Sorted_TopL}} \cap B_{\text{test}}|$ 为交集的节点数量, $|B_{\text{test}}|$ 指 B_{test} 集的节点数量。

2.4 间接关联 S_{AC} 的计算

目前已有多种指标用于评价 $A - B - C$ 者之间的关联性,笔者使用基于局部信息的节点相似性指标——共同邻居(Common Neighbors)^[11] 计算 $A - B - C$ 之间的潜在关联性 S_{AC} ,以辅助比较强弱关联在文献知识发现时的准确率 P_{weak} 与 P_{strong} ,如下所示:

$$S_{AC} = |\Gamma(A) \cap \Gamma(C)|$$

式中: $\Gamma(A)$ 与 $\Gamma(C)$ 分别为节点 A 与 C 的邻居,即两个节点如果有更多的共同邻居,则它们更倾向于形成连接。

3 数据测试

3.1 数据准备

以关键词“miRNA or MicroRNA”从 PubMed 中提取 51 118 条摘要数据,以 2012 年作为时间分割点,将 2002 - 2012 年的数据作为训练集,2013 - 2015 年的数据作为测试集,得到训练网络 G_1 包含节点 20 998,边 102 363;测试网络 G_2 包含节点 28 325,边 130 369。从训练集中随机选取 50 个关键词作为种子词项 A_{random} 。

3.2 结果与讨论

(1) 取 $L = 0.1, 0.2, 0.3, \dots, 1$,即取有序

结果集 C_{sorted} 前 10%、20%、10%、20%、30%、40%、50%、60%、70%、80%、90%、100% 的词项时, 计算准确率 P 的结果, 见图 2。(2) 缩小靠前区间, 取 $L=0.01, 0.02, 0.03, \dots, 0.1$, 即取有序结果集 C_{sorted} 前 1%、2%、3%、4%、5%、6%、7%、8%、9%、10% 的词项时, 计算准确率 P 的结果, 见图 3。

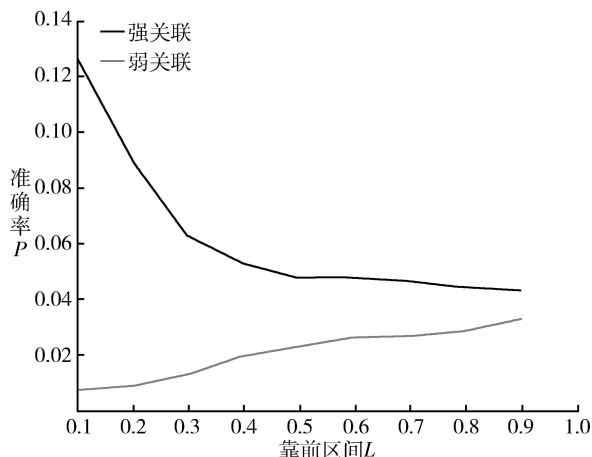


图2 $L=0.1, 0.2, 0.3, \dots, 1$ 强关联与弱关联的准确率 P 的比较

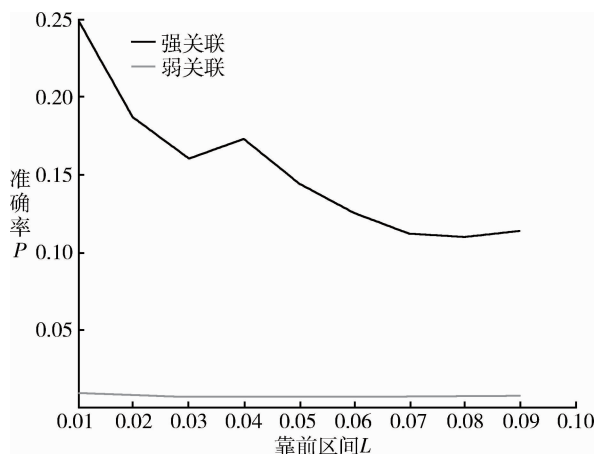


图3 $L=0.01, 0.02, 0.03, \dots, 0.1$ 强关联与弱关联的准确率 P 的比较

综合分析图 2、图 3 的测试结果可以发现, 由强关联计算得到的准确率 P_{strong} 显著高于由弱关联计算得到的准确率 P_{weak} , 说明利用强关联可以获得更多的新关联, 在短期内能衍生出较多的新研究内容。同时, 图 2 与图 3 的结果也表明利用关联性强弱对发现的间接关联进行排序, 强关联可以很好对

候选集实现置优排序, 表现出下降趋势; 弱关联总的趋势不明显, 且表现出上升趋势, 在靠前的区间准确率反而低于靠后的区间, 未达到置优排序的效果。尤其在图 2 中, 在最靠前的 $(0, 0.1]$ 区间中, 强关联计算得到的准确率显著好于其他区间, 下降趋势明显, 说明排序相对靠后的结果的表现不断降低, 符合一个文献知识发现系统的预期, 即在置优最前的区间越靠前的结果表现越优。从整体上来看, 测试数据的表现都未能获得较高的准确率, 这一方面是因为实体抽取时未过滤那些宽泛概念, 如 "cell"、"gene"、"protein" 等, 这些概念尽管邻接大量的关联, 但在具体关联发现过程中反而没有实际的意义; 另一方面, 简单的基于共现方法的关联抽取会给整个结果集带来很多假阳性关系, 增加了测试数据集的混杂性, 从而降低了总体的准确率。

面对科研工作中的实际研究问题, 文献知识发现的具体实施仍是一个启发式过程, 不管是强关联还是弱关联的评价方法, 统计计算所得到的间接关联都是需要进一步验证的假设, 但大量的文献集所蕴含的潜在关联的数量异常庞大, 通过计算发现帮助筛选相对有效的关系仍是目前一种可行的途径。间接关联性由于缺乏直接的语义关系的评价, 难以直接确证语义联系的存在, 但对于探索影响间接联系评价的因素与作用来说, 观察总体富集的效果, 不会影响 P_{strong} 和 P_{weak} 差异比较的表现, 因此, 强关联性可作为提升文献知识发现应用效果的重要因素。

4 结语

面对大数据时代知识获取的需求与挑战, 生物医学文献知识发现研究对完成从文献到知识的转化具有重要作用。本文分析了强关联和弱关联在生物医学文献知识发现过程的不同表现, 测试发现在知识发现过程中由强关联获得的准确率显著高于弱关联时的准确率, 表明利用强关联性可以获得更多有效的新关联, 在短期内能衍生出较多的新研究内容,

(下转第 20 页)

户的各类实体资源进行整合和信息资源共享,实现区域级医疗系统的构建,包括个人健康数据自动监控、智能化医疗诊断体系、自动化医保结算等子模块。(3) 在基于医疗服务流程再造的卫生服务模式中,患者可以在家搜集个人体征数据、及时反馈健康状况、与医护人员远程沟通、更加主动地参与各项治疗环节、使用便捷的结算系统和医疗保险报销系统等,期望实现“以患者为中心”的最终目标。(4) 在患者医疗数据和保险数据共享平台的构建时,涉及患者隐私以及支付安全,务必要确保数据平台的安全,下一步研究中也需从此角度入手保证信息安全。

参考文献

- 1 迈克尔·哈默,詹姆斯·钱皮. 企业再造: 企业革命的

(上接第 10 页)

强关联可作为提升生物医学文献知识发现系统的应用效果的重要因素。

在实际科研过程中文献知识发现系统面临的一个最大瓶颈是“噪声”关联过多导致的总体准确率不高,其主要原因是实体概念提取过程中存在很多宽泛的概念,而这些概念涉及的已有关联对于发现新的知识作用不大。当前科研人员为了提高文献知识发现系统的准确率,一方面通过语义限定对词项进行范围控制;另一方面不断改进系统的交互性,方便用户快速筛选与检索源文献,达到提升知识发现效率的目标。随着自然语言处理技术的发展,大规模文献的批量处理、语义层面的关系抽取将显著提升知识发现系统的准确率,生物医学文献知识发现系统应用也会越来越广泛。

参考文献

- 1 Swanson DR. Fish oil, Raynaud's Syndrome, and Undiscovered Public Knowledge [J]. *Perspect Biol Med*, 1986, 30 (1): 7-18.
- 2 刘俊丽,蒋勇青,张秀梅. 生物医学非相关文献知识发现研究综述 [J]. *情报科学*, 2014, 32 (1): 157-161.
- 3 Smalheiser NR, Torvik VI, Zhou W. Arrowsmith Two-node Search Interface: a tutorial on finding meaningful links be-

宣言书 [M]. 王珊珊,胡毓源,徐获洲译. 上海: 上海译文出版社, 2007.

- 2 李姝,王长波. 基于移动医疗技术的医疗流程再造分析 [J]. *内江科技*, 2015, (12): 63, 117.
- 3 陈晓萍,张涛,宗文红. 基于移动医疗技术的居民健康管理系统设计 [J]. *中国卫生信息管理杂志*, 2015, 12 (1): 25-28, 48.
- 4 许德俊,黄以宽,屈晓晖,等. 多级医疗卫生机构诊疗协同服务信息集成化研究 [J]. *中国卫生信息管理杂志*, 2016, 13 (3): 241-246.
- 5 何宇红,熊勇,尚武. 利用云计算打造智慧医疗管理模式 [J]. *医学信息学杂志*, 2016, 37 (4): 39-42, 62.
- 6 陆春吉,李亚子,任慧玲. 我国医疗保险信息化发展 [J]. *医学信息学杂志*, 2016, 37 (7): 57-61.

tween two disparate sets of articles in MEDLINE [J]. *Comput Methods Programs Biomed*, 2009, 94 (2): 190-197.

- 4 Tsuruoka Y, Miwa M, Hamamoto K, et al. Discovering and Visualizing Indirect Associations Between Biomedical Concepts [J]. *Bioinformatics*, 2011, 27 (13): 111-119.
- 5 Poon H, Quirk C, DeZiel C, et al. Literome: PubMed-scale genomic knowledge base in the cloud [J]. *Bioinformatics*, 2014, 30 (19): 2840-2842.
- 6 Fleuren WW, Verhoeven S, Frijters R, et al. CoPub update: CoPub 5.0 a text mining system to answer biological questions [J]. *Nucleic Acids Res*, 2011, (39): 450-454.
- 7 Yetisgen-Yildiz M, Pratt W. A New Evaluation Methodology for Literature-based Discovery systems [J]. *J Biomed Inform*, 2009, 42 (4): 633-643.
- 8 Weeber M, Kors JA, Mons B. Online Tools to Support Literature-based Discovery in the Life Sciences [J]. *Brief Bioinform*, 2005, 6 (3): 277-286.
- 9 闵波,刘爱中,郑萍,等. 基于复杂关联网络的生物医学研究结构的挖掘 [J]. *中华医学图书情报杂志*, 2015, 24 (8): 1-4.
- 10 Pavlopoulos GA, Promponas VJ, Ouzounis CA, et al. Biological Information Extraction and Co-occurrence Analysis [J]. *Methods Mol Biol*, 2014, (1159): 77-92.
- 11 吕琳媛. 复杂网络链路预测 [J]. *电子科技大学学报*, 2010, 39 (5): 651-661.