文章编号：2095-2163(2020)01-0233-04 中图分类号：TP391；Q811.4 文献标志码：A

基于本体的疾病关联搜索方法的研究

梅 祎，王亚东
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：随着医疗技术和生物科技的快速发展，生物领域的大数据急剧膨胀，数据的快速、有效检索成为了至关重要的问题。本文针对疾病本体及其相关数据的检索问题，提出了基于疾病本体的关联搜索算法。首先，根据多种医疗数据库中的原始数据，构建异构知识网络，之后，设计了在知识网络中进行关联搜索的算法，算法对节点与关键字之间的每条路径进行评分，并选取其中评分最高的路径作为该节点的最后得分，最终选取得分最高的若干个节点。结果表明，该算法有效地搜索出了与关键字关联度较大地数据。

关键词：疾病本体；搜索；异构网络

Research on disease association search algorithm based on ontology

MEI Yi, WANG Yadong
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 With the rapid development of medical technology and biotechnology, the biological data has been expanded rapidly, so that the rapid and efficient retrieval of data has become a crucial issue. In this paper, an association search algorithm based on disease ontology is proposed for the retrieval of disease ontology and its related data. First, the paper builds heterogeneous knowledge network based on a variety of medical databases. After that, the paper designs the association search algorithm in the knowledge network. The algorithm scores each path between the node and the keyword, and selects the path with the highest score as the final score of the node. Finally, the paper selects nodes with highest scores as the search results. The results show that the algorithm is effective in searching for data with large correlation with keywords.

【Key words】 disease ontology; search; heterogeneous network

0 引 言

近年来，随着医疗技术和生物科技的迅猛发展，生物领域的大数据急剧增加。其中，疾病本体和其它相关的数据如药物、基因及相关文献等是生物大数据中不可或缺的典型组成部分。随着数据量规模的日趋庞大，数据的快速、有效检索成为了至关重要的研究问题。

然而，主流的综合搜索引擎如百度、Google 等对疾病关键词的检索结果普遍来说都比较简单通俗，偏向科普的形式，适用于并不具备专业知识的普通用户，因而不能满足更高层面的研究需求，尤其是对专业的医疗工作者和研究者来说，搜索结果也尚未臻至专业和全面。

面向疾病领域的垂直搜索引擎不仅数目稀少，而且功能单一。例如，疾病浏览器 Disease Ontology (http://disease-ontology.org)，虽然搜索结果是有效的，但搜索结果未能按照相关性的大小给出有序展示，且包含的知识不够丰富，对用户来说，从其中提取自身想要的信息还需做进一步的筛选，是一个耗费时间和精力的过程。

基于此，本文提出了基于疾病本体的关联搜索，关联检索算法综合考虑了疾病本体及其相关数据之间的关联，为用户提供有效的检索信息，结果表明，关联搜索算法的搜索结果相关度高、内容丰富，此外，用户还可以根据个人的偏好，或者行为习惯，通过调整参数，修改搜索结果出现的先后顺序。

1 相关工作

搜索引擎主要分为 2 种：综合搜索引擎和面向专业领域的垂直搜索引擎[1-2]。对此可阐释解析如下。

(1)综合搜索引擎[3]。现在已经得到了广泛的研究。其中，Google 公司提出的 PageRank 算法[4]，即是依靠网页之间的链接关系来确定每一个页面的等级，一个页面到另一个页面的链接可解释成该页面对另一页面的投票，PageRank 算法根据这些投票来源和投票目的的等级确定新的等级。PageRank

作者简介：梅 祎(1993-)，女，硕士研究生，主要研究方向：生物信息学；王亚东(1964-)，男，教授，博士生导师，主要研究方向：机器学习、知识工程、生物信息学等。

收稿日期：2017-06-15

算法在普通的网页搜索中表现得较好,但是涉及到专业领域的搜索时,往往没有很好的结果。除了PageRank 之外,HITS 也是常用的链接分析技术[5],是用于分析网页重要性的设计算法,可根据一个网页的入度(指向此网页的超链接)和出度(从此网页指向别的网页)来衡量网页的重要性。其最直观的意义就是如果一个网页的重要性很高,则由其所指向的网页的重要性也会较高。一个重要的网页被另一个网页所指,则表明指向该网页的重要性也会很高。指向别的网页定义为 Hub 值,被指向定义为 Authority 值。

(2)垂直搜索引擎。即专业或专用搜索引擎,就是专为查询某一学科或主题的信息而产生的查询工具,专门收录一方面、某一行业或某一主题的信息,与搜索引擎门户相比,对解决实际查询问题要更加有效[6]。

2 方法

2.1 问题描述

在由不同类型的数据构成的数据集合中,数据之间存在着各种相关关系,比如疾病本体之间的父子及兄弟关系,疾病与药物、基因等数据之间的关联关系等。因此,这些数据组成了一个复杂的异构网络,每一个数据条目都可以对应成网络中的一个节点,数据间的关系对应节点间的边。

本课题所需要实现的目标是在这个复杂的异构网络中,搜索出与待搜索关键词相关度最高的前 n 个节点。

2.2 知识库建立

本课题首先建立起了疾病本体及其相关数据包括基因、表型、药物、文献等数据之间的异构网络,研究使用的数据主要来自于选取医学数据库,该部分研究内容可探讨分述如下。

(1)Disease ontology[7]。疾病本体来自于开放生物医学体系(Open Biomedical Ontology,OBO),是纪录了与人类相关疾病数据的本体库。

(2)MeSH(Medical Subject Headings)[8]。是美国国家医学图书馆创建的医学主题词表,由于其结构体系完整合理,在世界范围内被广泛使用。

(3)KEGG[9](Kyoto Encyclopedia of Genes and Genomes)数据库。是从分子水平,整合了基因、酶、化合物及代谢网络信息的综合性数据库。

(4)OMIM[10](Online Mendelian Inheritance in Man)数据库。是在线孟德尔人类遗传数据库,主要包括了遗传疾病、遗传表型及基因等信息。

(5)MEDIC(合成疾病词汇,merged disease vocabulary)[11]。是整合 OMIM 术语、同义词和标识符与 MeSH 术语、同义词、标识符、定义和层级关系的资源,通过 MEDIC,就在原有基础上补充及丰富了数据间的关联关系。

2.3 方法设计

知识库网络可以表示为图 $G=(V, E)$,其中 V 为节点集合,E 为边的集合,任意的 2 个节点之间用一条边连接,边的权值代表相似性或相关性,其取值范围为[0,1]。权值矩阵为 $\boldsymbol{M}$,对任意的节点 v_1, v_2,称 $\boldsymbol{M}(v_1, v_2)$ 为节点 v_1 和节点 v_2 之间的固有相关度。

衰减系数用来表示从某一类型节点到另一类型节点之间相关性的衰减。设节点类型数量为 N,则衰减系数可以表示为 $N\times N$ 的矩阵 $\boldsymbol{R}$。对 $\forall v\in V$,函数 $t(v)$ 表示节点 v 的节点类型。对任意的节点 v_1, v_2, $\boldsymbol{R}(t(v_1), t(v_2))$ 表示节点 v_1 和 v_2 的衰减系数。

综合考虑节点间的相关性和节点类型间的衰减系数,得到一个大小为 $|V|\times|V|$ 的综合相关度矩阵 $\boldsymbol{S}$。计算方法为:

$$\boldsymbol{S}_{i,j}=\boldsymbol{M}_{i,j}\times\boldsymbol{R}_{t(i),t(j)},\ i\in[1,|V|],\ j\in[1,|V|]. \tag{1}$$

对于待查询的关键字 k,可以将其视为一个特殊的节点,加入至上述网络。其与每个节点之间有一个直接的相似度,由指定的打分函数进行定义。将网络由此扩展后,新的网络可表示为一个$(|V|+1)\times(|V|+1)$ 的矩阵 $\boldsymbol{W}$。

查询算法的目的是找到网络中与待查询关键字 k 相关性最高的 n 个节点,而该相关性不仅由该节点与待查询关键字节点之间的直接相关性决定,亦是由该节点周围的节点决定。下面,研究给出了查询问题的形式化定义。

定义 1 设 v_i, v_{i+1}, …,v_{i+n} 为 v_i 到 v_{i+n} 之间的一条哈密顿通路,将 $\boldsymbol{W}(v_i, v_{i+1})\times\boldsymbol{W}(v_{i+1}, v_{i+2})\times\cdots\times\boldsymbol{W}(v_{i+n-1}, v_{i+n})$ 称为节点 v_i 和 v_{i+n} 之间关于该条哈密顿通路的相关度。若该条通路记为 p,将该路径相关度则记为 Y_p。

定义 2 设函数 $g(v_i, v_j)$ 表示节点 v_i 和节点 v_j 之间的全部哈密顿通路。节点 v_i 和节点 v_j 之间相关度可以定义为节点 v_i 和节点 v_j 之间的全部哈密顿通路的相关度的最大值,设节点间的相关度为 $f(v_i, v_j)$,则 $f(v_i, v_j)=\max\{Y_p\}$, $p\in g(v_i, v_j)$。当 v_i 为待搜索的关键词节点 k 时,记 $f(k, v_j)$ 为 $f(v_j)$。

因此，问题可以被表述为输入待查询关键词 k 和查询数目 n，输出集合 Q，满足 $Q \subset V$，$|Q| = n$，且 $\forall v' \in V - Q$，$\nexists v \in Q$，$f(k, v') > f(k, v)$（即输出分数前 n 的节点数组）。

在此基础上，研发得到了关联搜索算法的流程步骤可表述如下。

输入：查询的关键词 k，查询数目 n

输出：集合 Q，满足 $Q \subset V$，$|Q| = n$，且 $\forall v' \in V - Q$，$\nexists v \in Q$，$f(k, v') > f(k, v)$（即输出分数前 n 的节点数组）。

$A \leftarrow [\,]$

$B \leftarrow [\,]$

For v in V:

  $v.score \leftarrow \boldsymbol{W}(k, v)$

  $B.add(v)$

  $B.sort$

For i from 1 to n:

  $v \leftarrow B.first$

  $A.add(v)$

  $B.remove(v)$

  for v' in B:

  $v'.score = \max\{v'.score, v.score * \boldsymbol{W}(v, v')\}$

  $B.sort$

return A

2.4　算法正确性证明

考虑如下的命题：每次循环执行前，B 数组的第一个节点 v_1 满足 $f(v_1) = v_1.score$，且不存在 B 数组中的其它节点 v_i，会使 $f(v_i) > v_1.score$。下面，关于该命题的正确性证明过程详见如下。

经过研究可知，循环第一次执行前，显然，$v_i.score = \boldsymbol{W}(k, v_i)$，而 B 数组中元素已经经过排序，因此 $\boldsymbol{W}(k, v_i) < \boldsymbol{W}(k, v_1) = v_1.score$。设 v_i 与关键字节点 k 间的某条哈密顿通路 $p_i = <k, v_s, v_{s+1}, \cdots, v_j, \cdots, v_t, v_i>$ 是使得 Y_{p_i} 最大的一条通路，即 $f(v_i) = Y_{p_i} = \boldsymbol{W}(k, v_s) * \boldsymbol{W}(v_s, v_{s+1}) * \cdots * \boldsymbol{W}(v_t, v_i) \leqslant \boldsymbol{W}(k, v_s) \leqslant \boldsymbol{W}(k, v_1) = v_1.score$，所以 $f(v_i) \leqslant v_1.score$。同时 $f(v_1) = \max\{Y_p\}$，$p \in g(k, v_1)$，即使得 $f(v_1) \geqslant W(k, v_1) = v_1.score$。综上所述，$f(v_1) = v_1.score$。故而该命题在第一次循环执行前成立。

假设每次循环执行前，命题为真，则经过一次循环后，由于对数组 B 中的节点，都使用 $v_i.score = \max\{v_i.score, v_1.score * \boldsymbol{W}(v_1, v_i)\}$。设排序之后 B 数组中的第一个新节点为 v'_1，则有 $v'_1.score \geqslant v_i.score$。考虑 v_i 的所有与节点 k 之间的哈密顿通路。可将其分为如下 2 种类型。

（1）设某条通路为 p，且除了 k 节点和 v_i 节点外，其它节点都是 A 中的元素。此时，显然有 $Y_p \leqslant v_i.score \leqslant v'_1.score$；

（2）设某条路径为 p，除 k 节点和 v_i 节点外，其某些节点为 B 中的元素，不妨设第一次出现的 B 中的节点为 v_s。则 $Y_p \leqslant Y_{(k,\cdots,v_s)} \leqslant v_s.score \leqslant v'_1.score$。继而推得 $f(v_i) \leqslant v'_i.score$，而 $f(v'_1) \geqslant v'_1.score$，因此 $f(v'_1) = v'_1.score$。综上所述，当循环执行后，命题依然成立。

由于该命题成立，使得算法每次从 B 数组中移除第一个节点，将其添加到 A 数组中，而当循环执行 n 次结束后，A 数组中的节点是全部节点中分数从大到小排列的前 n 个元素。因此推得，算法是正确的。

2.5　算法复杂度分析

（1）算法的时间复杂度为：$n \times |V| \times \log(|V|)$。

（2）算法的空间复杂度为：$|V|^2$。

3　结果分析

以搜索“brain cancer”为例，可得搜索结果如图 1 所示。

图 1　“brain cancer”的搜索结果图

Fig. 1　The search results of keyword “brain cancer”

从图 1 中可以看出，首先搜索得到与搜索关键词直接相关的疾病本体“DOID：1319”和“DOID：4203”。然后搜索出了其它关联的疾病本体，如“DOID：3620”，从路径中可以看出其是与疾病本体“DOID：1319”相关的，这 2 个疾病本体在疾病本体树中的结构如图 2 所示，分析得知“DOID：1319”是疾病本体“DOID：3620”的子节点。图 3 是计算后的疾病本体间相似度，由此推得，这 2 个疾病本体具有较高的相关度。

接下来，若以“0050120”作为搜索关键词，搜索的结果则如图 4 所示。从图 4 中可以看出，不仅搜索出了直接与关键词匹配的疾病本体“DOID：0050120”，还检索出了该疾病本体相关的基因及表型等。此外，研究中又发现，通过疾病本体“DOID：

0050120”关联到了的基因“5515”再关联到疾病本体“DOID:0060060”,体现了这 2 个疾病本体关联着共同的基因,两者之间也存在着某种关联关系。

图 2 疾病本体“DOID:1319”与疾病本体“DOID:3620”在树中的位置图

Fig. 2 The “DOID:1319” and “DOID:3620” in the tree of disease ontology

图 3 疾病本体“DOID:1319”与其它疾病本体的相似度信息图

Fig. 3 The similarity of“DOID:1319” with other disease ontologies

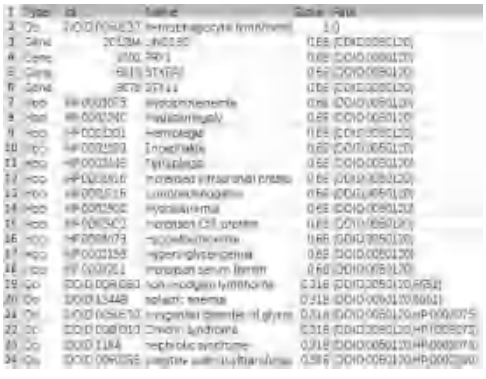

图 4 “0050120”的搜索结果图

Fig. 4 The search result of keyword “0050120”

4 结束语

本文针对疾病本体及其相关数据的检索问题,提出了基于疾病本体的关联搜索算法。首先,根据多种医疗数据库中的原始数据,构建异构知识网络,之后,设计了在知识网络中进行关联搜索的算法,算法可对节点与关键字之间的每条路径进行评分,并选取其中评分最高的路径作为该节点的最后得分,最终选取得分最高的若干个节点。结果表明,关联搜索算法不仅可以搜索出文本匹配方法能够搜索出的结果,而且能够搜索出其难于搜索得出、但却与关键词相关的结果。

参考文献

[1] ALMPANIDIS G, KOTROPOULOS C, PITAS I. Combining text and link analysis for focused crawling—An application for vertical search engines[J]. Information Systems, 2007, 32(6):886-908.

[2] 刘畅. 综合搜索引擎与垂直搜索引擎的比较研究[J]. 情报科学, 2007, 25(1):97-102.

[3] BRIN S, PAGE L. Reprint of: The anatomy of a large-scale hypertextual web search engine[J]. Computer Networks, 2012, 56(18):3825-3833.

[4] HAVELIWALA T H. Topic-sensitive PageRank [C]// Proceedings of the 11th International Word Wide Web Conference (WWW 2002). Hawai:ACM, 2002:517-526.

[5] YAN Lili, WEI Yingbin, GUI Zhanji, et al. Research on PageRank and hyperlink-induced topic search in Web structure mining [C]// 2011 International Conference on Internet Technology and Applications.Wuhan, China: IEEE, 2011:1-4.

[6] NIE Z, WEN J R, MA W Y. Object-level vertical search[C]// Third Biennial Conference on Innovative Data Systems Research (CIDR 2007). Asilomar, CA, USA: Dblp, 2007:235-246.

[7] RESNIK P. Using information content to evaluate semantic similarity in a taxonomy[C]// IJCAI'95 Proceedings of the 14th International Joint Conference on Artificial Intelligence-Volume 1. Montreal, Quebec, Canada:ACM,2015:448-453.

[8] ROGERS F B. Medical subject headings[J]. Bulletin of the Medical Library Association, 1963, 51(2):114.

[9] RÉDEI G P. Kyoto encyclopedia of genes and genomes[M]// Encyclopedia of genetics, genomics, proteomics and Informatics. Dordrecht: Springer, 2008:9130.

[10] HAMOSH A, SCOTT A F, AMBERGER J, et al. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders[J]. Nucleic Acids Research, 2005, 33(1):514-517.

[11] DAVIS A P, WIEGERS T C, ROSENSTEIN M C, et al. MEDIC: A practical disease vocabulary used at the Comparative Toxicogenomics Database[J]. Database : the journal of biological databases and curation, 2012, 2012(2012):bar065.

(上接第 232 页)

[3] OREL S G, SCHNALL M D. MR imaging of the breast for the detection, diagnosis,and staging of breast cancer1[J]. Radiology, 2001,220(1):13-30.

[4] 何锡嘉,凌巍高,张雅欣,等. 数字化医学影像技术下多模态图像配准仿真[J]. 计算机仿真,2018,35(12):166-170.

[5] 李胜东,吕学强. 基于图片问答的静态重启随机梯度下降算法[J]. 计算机研究与发展,2019,56(5):1092-1100.

[6] MAHEEN A S S, ZAIN B M, ALI L A, et al. Accuracy of apparent diffusion coefficients and enhancement ratios on magnetic resonance imaging in differentiating primary cerebral lymphomas from glioblastoma[J]. The neuroradiology Journal,2019.

[7] 李晶辉. 基于互信息的多层隐朴素贝叶斯算法研究[D]. 长沙:湖南大学,2012.

[8] 赵博文,王灵矫,郭华. 基于泊松分布的加权朴素贝叶斯文本分类算法[J/OL]. 计算机工程:1-7[2019-05-27]. https://doi.org/10.19678/j.issn.1000-3428.0054056.