·开发研究与工程应用· 文章编号:1000-3428(2015)01-0284-05 文献标识码:A 中图分类号:TP311

基于语义关系的疾病知识提取系统

吴晓芳,杨志豪,林鸿飞,王 健

(大连理工大学计算机科学与技术学院,辽宁 大连 116024)

摘 要:在生物医学领域,通过知识提取过程从海量的生物医学文献中提取疾病、基因和药物之间的关系并可视化显示,可以为临床医学实验提供有效的假设检验,推动生物医学科技的发展。为此,提出一种基于语义关系的以疾病为中心的疾病、基因和药物间的知识提取系统。利用 SemRep 得到特定主题 Medline 文献的语义输出,通过显著信息提取算法提取 SemRep 的语义输出关系。对照 OMIM 和 GHR 在线数据库进行评估,实验结果显示该显著信息提取系统的准确率可达 0.76。

关键词:知识提取;语义关系提取;显著信息提取算法;SemRep 工具;语义输出;网络图可视化

中文引用格式:吴晓芳,杨志豪,林鸿飞,等. 基于语义关系的疾病知识提取系统[J]. 计算机工程,2015,41(1):284-288.

英文引用格式:Wu Xiaofang,Yang Zhihao,Lin Hongfei,et al. Disease Knowledge Extraction System Based on Semantic Relation[J]. Computer Engineering,2015,41(1):284-288.

Disease Knowledge Extraction System Based on Semantic Relation

WU Xiaofang,YANG Zhihao,LIN Hongfei,WANG Jian

(School of Computer Science and Technology,Dalian University of Technology,Dalian 116024,China)

【Abstract】 In the biomedical field,knowledge summarization can greatly promote the innovation of biomedical science and technology. Dynamic summarization can provide novel clinical experimental hypothesis by extracting the links among diseases,genes,drugs from the mass of biomedical literature and visualizing it. This paper presents a system which summarizes the salient relations by the salient extraction algorithm using the specific subject Medline corpus by SemRep semantic output. Experimental results show that the precise of experimental result is 0.76 referring to OMIM and GHR online databases.

【Key words】 knowledge extraction;semantic relation extraction;significant information extraction algorithm;SemRep tool;semantic output;network diagram visualization

DOI:10.3969/j.issn.1000-3428.2015.01.054

1 概述

生物医学文献持续不断的增长给传统的信息检索技术带来极大的挑战。有效的医学文献检索,尤其是从海量的生物医学文献中发现显著的疾病、基因、药物之间的关联信息对生物医学工作者在临床试验和病患诊疗方面有极大的帮助。传统的人工阅读大量文献费时费力且效果甚微,在如今的数字化信息时代已经不再适用。虽然早先的信息检索技术已经应用到生物医学领域的知识提取,但是信息检索算法的有效性并没有得到很好的评估[1]。文献[2]从 Medline 中抽取有用的关系,简洁地概括出原始文献的主要信息。文献[3]提出了一个自动从文献集中提取摘要的算法 Combo,该算法提取了与某一特定疾病相关的基因。文献[4]在之前实验基础上又提出了一个用于决策支持的文本摘要生成方法。为了跟踪最新的医学研究领域的工作进展,生物医学文献的高效检索,有效关系的提取和展示对临床决策支持[5]来说尤为重要。自动摘要方法[6]在信息提取中有较好的效果,但是自动摘要最终形成

基金项目:国家自然科学基金资助项目(61070098,61272373,61340020);中央高校基本科研业务费专项基金资助项目(DUT13JB09);国家社会科学基金资助项目(08BTQ025)。

作者简介:吴晓芳(1989-),女,硕士研究生,主研方向:知识发现,文本挖掘;杨志豪,副教授、博士、博士生导师;林鸿飞,教授、博士、博士生导师;王 健,副教授。

收稿日期:2013-12-30 **修回日期**:2014-03-14 **E-mail**:xfwu@mail.dlut.edu.cn

的依然是文本形式,不够直观。因此,需要有效基于语义关系抽取的算法来从大量的生物医学文献中提取出重要的实体关联信息,并用可视化的方法将该关联信息呈现给医学工作者。

与以往研究不同,本文提出一个基于语义关系的以疾病为中心的疾病、基因和药物间的知识提取系统。该系统利用从 Medline 生物医学数据库检索到的相关疾病的语料集,运用 SemRep 工具处理得到相关疾病语料集的语义输出。通过显著信息提取算法筛选出以疾病为中心的疾病、基因和药物三者之间重要的关联信息,并以网络关系图的形式呈现给生物医学工作者。

2　系统方法

2.1　系统流程

系统流程如图 1 所示。

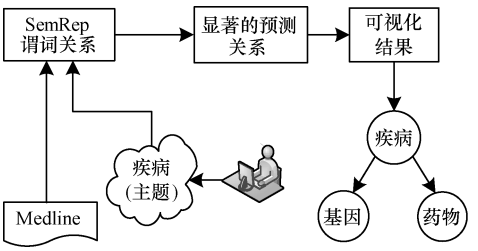

图 1　系统流程

对于特定的疾病,从 PubMed 上检索到 2003 年-2013 年与疾病相关的文献集。针对疾病和基因、疾病和药物给出不同的检索语句,检索得到相应的文献集。

通过 SemRep 工具处理文献集得到相应的语义输出。SemRep 能够从 Medline 语料的句子中抽取出 2 个实体之间的关联关系。如果一个句子中存在多个实体词和关系连接词,那么 SemRep 通过算法给每个关系打分,取分数最高的连接关系作为语义输出。

用 KL 散度、RlogF 矩阵显著信息评价算法分别对谓词关系、谓词关系连接的实体语义类型进行筛选,利用 PredScal 平衡前 2 种算法间的数值差,综合 3 种算法共同完成对疾病和基因、疾病和药物显著信息的提取。

最后将提取得到的以疾病为中心的显著信息网络图可视化,在系统界面中呈现给用户。

2.2　文献语料处理工具 SemRep

SemRep[7] 是一个基于规则自动从文献中识别关系预测的自然语言处理系统。SemRep 集成了 MetaMap 规范化的概念实体,并通过谓词关系将不同的实体概念连接起来。此外,SemRep 为每个实体词定义了相关的语义类型,方便特征选取和语义类型过滤。SemRep 提取的关系是根据 UMLS 的规则进行输出的,其原始结果中包含有很多条目,主要用到其中的实体名、语义类型和谓词关系部分。

例如,对于句子:

Expression levels of CBX7 inversely correlate with the progression of tumor stage and grade in urothelial carcinomas of the bladder, suggesting that downregulation of CBX7 indicates aggressive urothelial carcinoma phenotype.

SemRep 可以得到如下的语义输出:

SE|18984978|RESULTS|ab|5|relation|5|1|||gngm,aapp|gngm|23492|CBX7|CBX7||||1000|53|56|VERB|PART_OF||71|79|2|1|C0007138|Carcinoma,Transitional Cell|neop|neop|||urothelial carcinomas|||981|84|104

这里主要关注的是关联信息[8]:

CBX7| gngm|PART_OF| urothelial carcinomas| neop

CBX7 是一种参与调控细胞增殖衰老的转录抑制因子。从得到的输出可以看出,CBX7 转录抑制因子是癌细胞病变因子的组成部分。

关联信息是一个三元组(概念 1|语义类型,Predication,概念 2|语义类型)[9],概念 1 和概念 2 是 UMLS 的超级叙词表中定义的概念,每个概念包含该概念的标准化表示、概念标示符(Concept Unique Identifier,CUI)和语义类型。UMLS 的语义网络中共定义了 54 中谓词关系(PART_OF 是其中之一)。利用 SemRep 可以从一个句子中得到出一个或多个语义输出,通过一定的算法,对得到的语义输出进行打分,选取得分高的语义输出作为该句的关联信息。从文献中所有的句子里抽取出关联信息集,进一步运用显著信息提取算法进行筛选。

2.3　实验数据

以膀胱癌(Carcinoma of bladder)为例,介绍实验中用到的数据集以及显著信息提取算法的实现。

(1)与 Carcinoma of bladder 相关的基因方面的文献集 A

("2003/01/01"[Publication Date]:"2013/07/31"[Publication Date]) AND (Urinary Bladder Neoplasms/genetics[majr] AND Urinary Bladder Neoplasms/etiology[majr]) AND English[la]AND humans[mh]

(2)与 Carcinoma of bladder 相关的药物方面的文献集 B

("2003/01/01"[Publication Date]:"2013/07/31"[Publication Date]) AND Urinary Bladder Neoplasms[mh noexp] AND drug therapy[sh] AND Clinical Trial[pt] AND English[Lang] AND humans[mh]

这 2 组查询语句检索了从 2003 年-2013 年的 Medline 文献。与基因相关的文献集 A 设定了基因和膀胱病因学等限制词,检索得到与膀胱癌相关的基因类的文献。与药物相关的文献集 B 设定了药物、临床治疗和膀胱病因学等限制词,检索得到与膀胱癌相关的药物类文献。通过上面 2 组查询语句,从 PubMed 上下载对应的 Medline 文献集。

2.4 显著信息评价算法

为了实现有用信息的提取,本文实验中使用了 3 种显著信息提取算法,自动地从 SemRep 的输出结果中筛选出查询的疾病与基因、药物之间的关联关系,排除掉繁多的相关性弱的关系。这 3 种显著信息提取算法介绍如下:

(1)KL 散度

KL 散度[10],又叫相对熵,在信息论中用于衡量 2 个概率分布的相对距离。在这里对关系谓词在疾病数据集 A 中的概率 P 和关系谓词在所有数据集 B 中的概率 Q 作为要衡量的 2 个概率。相对距离大的关系谓词表示在该疾病数据集中有比较突出的作用,从而通过得到的 KLD(Kullback-Leibler Divergence)得分值对关系谓词进行排名,得到关系谓词的筛选结果。

$$D(P||Q)=\sum P(x)\mathrm{lb}(P(x)/Q(x))$$

其中,x 代表一个关系谓词;$P(x)$ 代表关系谓词 x 在分布 P 中的概率;$Q(x)$ 代表关系谓词 x 在分布 Q 中的概率。例如,关系谓词 ASSOCIATED_WITH 在分布 P 中的概率为 0.290,在分布 Q 中的概率为 0.076,那么关系谓词 ASSOCIATED_WITH 的 KLD 值为 0.560 3。

KLD 算法中分布 Q 的统计数据选取了 2003 年 1 月 1 日-2013 年 7 月 31 日之间所有的 Medline 文献集。

(2)RlogF

RlogF 矩阵[11]旨在得到 SemRep 输出中同一个关系谓词相关度较高的语义类型,用函数 R 表示。关系谓词在做统计的时候受限于它在 SemRep 中的语义类型。

把检索词 Carcinoma of bladder 的语义类型 neop 作为种子语义类型。因为数据集是跟 Carcinoma of bladder 直接相关的,所以得到的语义类型中定有很多的 neop,排除掉该语义类型的影响,从而能更好地筛选出与该语义类型相关联的非种子语义类型。

$$R(pattern_i)=\mathrm{lb}(semantic\ type\ frequency_i)\times P(relevant|pattern_i)$$

其中,条件概率($P(relevant|pattern_i)$)是在语料 A 中出现的与关系谓词直接相关的实体的语义类型的个数(包含重复的部分)与所有出现的语义类型个数的比例。

$$P(relevant|pattern_i)=\frac{semantic\ type\ frequency_i}{total\ frequency_i}$$

例如,如果与关系谓词 ASSOCIATED_WITH 共现的非种子语义类型 gngm 在文献集 A 中出现 107 次,所有与关系谓词 ASSOCIATED_WITH 共现的非种子语义类型共有 171 个(包含重复的部分),那么关系谓词 ASSOCIATED_WITH 的 RlogF 值为 4.22。

(3)PredScal

RlogF 算法得到的值会远远超过 KLD 算法得到的值,在衡量一个关系的时候,RlogF 的结果占很大的比例。为了共同引用 2 种算法的思想,引入一个尺度函数 p 作为平衡因子来调整 2 个函数在同一数据集中的计算结果。

$$p=1/\mathrm{lb}(c)$$

在这个计算中,c 代表不同的关系谓词个数。例如,如果数据集中有 16 个不同的关系谓词,那么 PredScal 的平衡因子值 0.25。

以上 3 种算法结合起来共同完成对 SemRep 的输出结果的信息提取,用 Summa 算法来表示每个谓词关系的分值,运算结果表示为 *summa*。

$$summa=D\times R\times p$$

对于 SemRep 输出中的每一个关系,将谓词和语义类型分别通过算法 KLD 和 RlogF 筛选出来,通过算法 PredScal 来矫正 2 个结果数值间的成倍差距。这样每个关系都有一个 Summa 值来作为它们的显著程度的量化。

3 实验与评价

实验中基因部分的信息提取共得到与疾病 Carcinoma of bladder 相关的基因 54 个。参照 Online Mendelian Inheritance in Man(OMIM)和 Genetics Home Reference(GHR)中的基因文献记录进行标注,在得到的 54 个基因中有 41 个与疾病 Carcinoma of bladder 在 OMIM 和 GHR 里有关联关系。由此计算得出,实验提取结果的准确率为 0.76。而 SemRep 语料中抽取的实体之间的关系准确率为 0.73,召回率为 0.55,综合分类率 F 值为 0.63[12],本文显著信息提取算法的准确率有所提升。

3.1 疾病与基因的关系

运用 KLD 算法得到了与 Carcinoma of bladder 相关的关系谓词,通过对关系谓词排序筛选出前5 个得分最高的实验结果,见表 1。从表 1 可以看出关系谓词 ASSOCIATED_WITH 得分最高,这说明在疾病 Carcinoma of bladder 与基因的关系中,它们之间的相互作用关系,由 ASSOCIATED_WITH 关系词所连

接的关系尤其重要。生物医学工作者可以从这个关系中寻找到与该疾病相关的基因,从而更有效地找到治疗该疾病的基因方法。

表 1　KLD 算法得到的前 5 个关系谓词 (与基因相关)

关系谓词	前 5 个得分
ASSOCIATED_WITH	0.258 4
PART_OF	0.215 7
COEXISTS_WITH	0.062 3
PREDISPOSES	0.037 9
AFFECTS	0.013 2

运用 RlogF 算法得到了与 Carcinoma of bladder 相关的谓词以及语义类型之间的关系排名,筛选出前 5 个得分最高的实验结果,见表 2。从表 2 可以看出语义类型 gngm 与关系谓词 ASSOCIATED_WITH 得分最高,这说明在疾病 Carcinoma of bladder 与基因之间的相互作用关系中,由谓词 ASSOCIATED_WITH 所连接的实体类型为 gngm 的关系最为突出。语义类型 gngm 是 Gene or Genome 的缩写,代表基因类。从结果中可以看出,运用显著信息提取算法有效地筛选出了跟疾病相关的基因。

表 2　RlogF 算法得到的数据(与基因相关)

语义类型	关系谓词	前 5 个得分
gngm	ASSOCIATED_WITH	3.546 9
humn	LOCATION_OF	3.459 4
phsu	TREATS	2.456 4
aapp	ASSOCIATED_WITH	2.000 6
gngm	PART_OF	1.212 4

以上 2 种算法,用 PredScal 算法做权衡后,得到疾病与基因相关的 Summa 的排名结果,见表 3。

表 3　Summa 信息提取的前 5 个结果(与基因相关)

语义类型	关系谓词	前 5 个得分
tisu	PART_OF	0.355 1
gngm	ASSOCIATED_WITH	0.212 1
celc	PART_OF	0.208 4
aapp	ASSOCIATED_WITH	0.119 6
cell	PART_OF	0.098 5

3.2　疾病与药物的关系

运用 KLD 算法得到了与 Carcinoma of bladder 相关的谓词,通过对关系谓词排序筛选出前 5 个得分最高的实验结果,见表 4。从表中 4 可以看出关系谓词 TREATS 得分最高,这说明在疾病 Carcinoma of bladder 与药物之间的相互作用中,由谓词 TREATS 所连接的关系尤其重要,通过 KLD 算法有效地找到了治疗疾病的相关药物。

表 4　KLD 算法得到的前 5 个关系谓词(与药物相关)

关系谓词	前 5 个得分
TREATS	0.680 8
PART_OF	0.044 7
USES	0.031 0
PRECEDES	0.022 5
PROCESS_OF	0.020 9

运用 RlogF 算法得到了与 Carcinoma of bladder 相关的谓词以及语义类型之间的关系排名,筛选出前 5 个得分最高的实验结果,如表 5 所示。从表 5 中可以看出语义类型 phsu 与关系谓词 TREATS 得分最高。这说明,在疾病与基因之间的相互作用关系中,由谓词 TREATS 所连接的实体类型为 phsu 的关系最为突出。语义类型 phsu 是 Pharmacologic Substance 的缩写,代表药物学物质。结果表明,显著信息提取算法有效地筛选出了能治疗疾病 Carcinoma of bladder 的药物。

表 5　RlogF 算法得到的数据(与药物相关)

语义类型	关系谓词	前 5 个得分
phsu	TREATS	8.043 9
fndg	PROCESS_OF	5.585 0
dsyn	PROCESS_OF	5.039 2
topp	TREATS	4.687 3
cell	LOCATION_OF	3.924 3

以上 2 种算法,用 PredScal 算法做权衡后,得到疾病与药物相关的 Summa 的排名结果,如表 6 所示。

表 6　Summa 信息提取的前 5 个结果(与药物相关)

语义类型	关系谓词	前 5 个得分
phsu	TREATS	1.727 6
antb	TREATS	0.905 7
topp	TREATS	0.862 4
medd	TREATS	0.429 5
hlca	TREATS	0.390 9

3.3　基因与药物的关系

通过 Summa 算法得到了疾病 Carcinoma of bladder 分别与基因、药物的相关关系实体集合。对得到的基因和疾病词对依次在 SemRep 数据库中进行检索,得到了基因和药物之间的关联关系。表 7 为选取的部分相关的基因和药物。

表 7　部分基因和药物的关联关系

基因	预测	药物
Tumor Suppressor Genes	STIMULATES	Cisplatin
FANCF	INTERACTS_WITH	Cisplatin
plakoglobin	STIMULATES	Cisplatin
Cadherins	INTERACTS_WITH	Cisplatin

4 系统描述

4.1 JUNG 工具包

系统可视化用到的 JUNG[13]（Java Universal Network/Graph framework）是一个 Java 开源项目，其目的在于为开发关于图或网络结构的应用程序提供一个易用、通用的基础架构。在系统实现过程中，使用 JUNG 功能调用，可以方便地构造图或网络的数据结构。应用经典算法如聚类、最短路径、最大流量等，编写和测试用户自己的算法，以及可视化的显示数据的网络图。

4.2 系统界面

图 2 中的网络图是以疾病 Carcinoma of bladder 为中心的疾病和基因、药物的关联信息。

浅色的结点表示的是跟疾病相关的基因，深色的结点表示的是跟疾病相关的药物。同时，部分基因和药物的关联关系也在图中展示出。

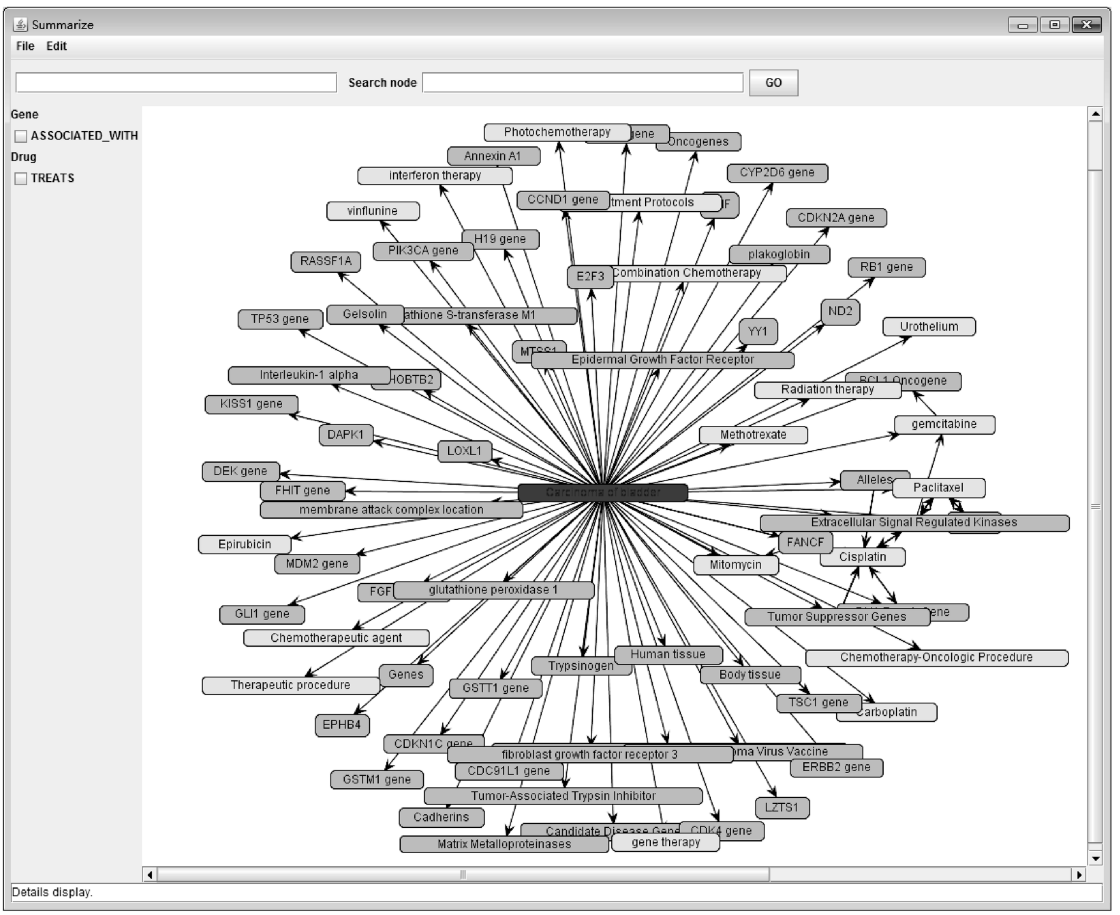

图 2 系统初始化显示及结点详细信息显示

对于整个网络图，编辑栏可以选择整体移动（TRANSFORMING）和部分选取（PICKING）功能。在选择（PICKING）功能，选择图中的任何一个结点，在底部面板的 Details display 栏显示该结点的详细信息，包括实体所在的 PubMed 文档号和包含该实体的句子。在 Search node 搜索框，输入一个疾病，便可手动检索疾病，并将该结点移至面板中心，在底部显示该结点的详细信息。左边的复选框用于单独显示某个模块、关系的单独子图。例如，选择 Gene 模块中的 ASSOCIATED_WITH 就可以单独显示与疾病相关的基因，这些基因跟疾病之间的谓词关系为 ASSOCIATED_WITH。单个关系的子图可以更方便用户找到与疾病有显著关系的基因和药物，有针对性地对得到的关系进行分析，提高生物医学工作者的查询效率。

5 结束语

信息提取在生物医学领域发展迅速，信息时代的科技发展需要高效的工具作为辅助。本文在提出信息提取算法的基础上，以疾病为中心，将疾病、基因和药物三者信息集成在可视化系统中。该系统有利于医学工作者快速了解跟疾病相关的基因信息，并能根据得到的药物信息对病情进行有效的分析和诊断。在算法方面，结果的准确率还有欠缺，下一步将研究改进方向并应用到信息提取中，完善系统功能。

（下转第 295 页）