· 生物医药信息研究

文本挖掘在生物医学领域中的应用及其系统工具

吕 婷[1], 姜友好[2]

[摘要] 系统介绍了生物医学文本挖掘的具体流程和文本挖掘技术在生物医学领域中的应用情况,并着重从自然语言处理和本体、命名实体识别、关系抽取、文本分类与聚类、共现分析、系统工具及评价、可视化等方面分别做了阐述。

[关键词] 生物医学文本挖掘; 自然语言处理; 命名实体识别; 关系抽取; 共现分析

[中图分类号] R318; G254.0 文献标识码: A [文章编号] 1671 - 3982(2010)04 - 0056 - 09

Application of text mining in biomedical field and its system tools

LU Ting[1], Jiang You - hao[2]

(1. Medical Library of Chinese PLA, Beijing 100039, China;
2. Department of Medical Information, Zhongnan University, Changsha 410013, Hunan Province, China)

[**Abstract**] the specific processes of text mining in biomedicine and the application of text mining technology in biomedical field were introduced in detail with stress laid on the natural language processing, ontology, named entity recognition, relationship extraction, text classification and clustering, co - occurrence analysis, system tools and their evaluation, and visualization.

[**Key Words**] text mining in biomedicine; natural language processing; named entity recognition; relationship extraction; co - occurrence analysis

1 文本挖掘概述

1.1 概念

数据挖掘(Data mining),又称数据库知识发现(Knowledge discovery in database),是指从结构化信息中提取人们感兴趣的知识。这些知识是隐含的、事先未知的、潜在的有用信息。文本挖掘(Text mining)是数据挖掘的一个方向,它所挖掘的对象是非结构化或半结构化,即从数以百万计的文本数据中寻找潜在规律和趋势。文本挖掘在商业、传媒、教育、政府、银行及生物技术、医疗卫生等行业领域都发挥着不可忽视的作用[1]。搜索引擎、自动邮件回复、垃圾邮件过滤、客户关系管理、自动简历评审等都是典型的文本挖掘技术。

1.2 流程及模型

文本挖掘的基本思想是利用文本切分技术抽取文本特征,将文本数据转化为计算机能识别的结构化数据,然后利用聚类、分类等数据挖掘技术形成结构化文本,并根据该结构发现新的概念及获取相应的关系。构成模型如图 1 所示。

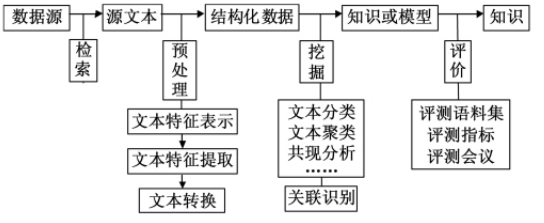

图 1 文本挖掘基本模型

1.3 技术

文本挖掘涉及多个学科领域,如数据库、信息检索、信息提取、机器学习、自然语言处理、计算语言学、统计数据分析、图论等。文本挖掘按照挖掘对象分为两类。一是单文本的数据挖掘,主要涉及的挖掘技术有文本摘要、信息提取(包括名字提取、短语提取和关系提取等)。二是文本集的数据挖掘,主要技术有文本分类、文本聚类、个性化文本过

[作者单位] 1. 解放军医学图书馆,北京 100039; 2. 中南大学医学信息系,湖南 长沙 410013

[作者简介] 吕 婷(1985 -),女,陕西宝鸡人,本科,发表论文 5 篇。

滤、文档作者归属、因素分析等。

以“预处理”过程为例，需要对文本数据做以下预处理：消除噪声和冗余数据，推算缺失数据，数据缩减，对元数据进行标记，词性标记，短语边界辨认，对特征项量化处理等[2]，最后形成计算机可处理的结构化信息。

2　生物医学文本挖掘

几个世纪以来，虽然科学信息都呈指数级增长，但现代医学文献数量之多仍让人印象深刻[3]。遗憾的是，人们对信息处理及分析的速度远远落后于信息本身的增长，从而产生了信息过载的问题[4]。生物信息文本挖掘就是通过计算机，帮助人们从爆炸式增长的生物医学自然语言文本数据中发现知识、抽取关系，减轻研究人员信息超载的负担。总的来说，生物医学文本挖掘可以从文献中抽取出特定的事实信息（主要是生物实体如基因、蛋白质、药物、疾病之间的关系），对整个生物知识网络的建立、生物体关系的预测、新药的研制等均具有重要的意义[5]。

2.1　自然语言处理与本体

2.1.1　自然语言的模糊性

计算语言学的研究使人们更关注对语言的理解分析。自然语言的模糊性使找到句子含义变得复杂，常常会出现不同的理解，如词汇歧义、句法歧义、语义歧义等。词汇歧义也称词类歧义和类别歧义，主要是因为一个单词可能有不止一种词性。自然语言的文法通常是模棱两可的，这就出现了句法歧义，如“AFB1 binds preferentially to DNA with an alternating G－C sequence compared to DNA with a sequence of contiguous Gs or Cs”与“GMPPCP binds to tubulin with a low affinity relative to GTP or GDP”。第一句中的“with”引出的介词短语修饰前面的“DNA”，第二句中的介词短语则是修饰“bind”前的“GMPPCP”，而并非“tubulin”。因为一个句子通常可能有多棵剖析树（Parse Tree），只有依靠语意及前后文意思，才能在其中选择一棵最适合的树。语义歧义涉及句子意思解释的问题，单词有不同含义时就会出现，如 figure 在“figure indicate”中指的是数字，而在“a good figure”中则指的是身材[6]。所以，要真正理解人类语言，需要有广泛的知识并要结合语境，而不是仅了解语言本身。

2.1.2　自然语言处理的应用

自然语言处理（Natural language processing，NLP）是人工智能（Artificial lntelligence，AI）和语言学领域的分支学科，主要用于中文自动分词（Chinese word segmentation）、词性标注（Part－of－speech tagging）、句法分析（Parsing）、自然语言生成（Natural language generation）、文本分类（Text categorization）、信息检索（Information retrieval）、信息抽取（Information extraction）、问答系统（Question answering）、机器翻译（Machine translation）、自动摘要（Automatic summarization）等。自然语言处理也可用于临床决策支持。如 Joshua Denny 等[7]调查发现，利用 NLP 和正则表达式查询心脏病专家对心电图的解释，可以更有效地识别 QTc 延长和其他心电图异常报告。

2.1.3　基于自然语言处理技术的文本挖掘系统

基于自然语言处理技术的文本挖掘系统有 MetaMap，IndexFinder，MedScan[8]，GeneWays[9－10]，PASTA[11]等。MedLEE 系统提取 UMLS 概念的查全率和查准率已分别达到 83% 和 89%[12]。P. Karina Tulipano 等[13]将 BioMedLEE 系统应用于分子成像领域，使用自然语言处理技术，通过结构化自由文本，找到相关的图像说明和文献，以协助自动标引和组织图像。如果没有一种方法能组织这些图像，很难完成图像的比较研究，成像技术解决方案如基于内容的图像检索将受到限制。结果此次 BioMedLEE 的查全率和查准率达到了 0.74 和 0.70。袁毅等[14]称其开发的基因相关文献挖掘网络平台是我国唯一基于自然语言处理的文本挖掘系统，能够通过文献获取、语法处理、语义处理、信息整合及可视化等步骤实现基因功能、基因与疾病关系、生物分子相互作用网络知识发现，辅助形成生物科学研究创新假设，准确率达 86%。

2.1.4　本体

在生物医学领域中，本体（Ontology）已经广泛用于领域专业知识的结构化组织。本体是对概念体系的明确的、形式化、可共享的规范说明。大量面向医学的本体被集成在一体化医学语言系统（UMLS）中[15]。

2.1.4.1　UMLS

一体化医学语言系统（UMLS）是对生物医学科学领域内许多受控词表的一部纲目式汇编，收录了 100 多部受控词表和分类系统，如 ICD－9－CM，ICD－10，MeSH，SNOMED CT，LOINC，世界卫生组织药物

不良反应术语集(WHO Adverse Drug Reaction Terminology, WHO - ART)、英国临床术语(UK Clinical Terms,又称为 Read Codes)、RxNORM、基因本体(Gene Ontology,GO)和 OMIM 等。它提供一种映射结构,使这些不同的术语系统之间能够相互转化[16]。UMLS 主要由超级叙词表(Metathesaurus)、语义网络(Semantic Network)、专家词典(Spseclalist Lexicon)构成[15]。相关支持性软件包括 MetamorphoSys[17-18], Lexical Variant Generation (LVG)[19]和 MetaMap[20]等。

2.1.4.2 基因本体

基因本体 GO[21]是 2000 年基因本体联盟(The Gene Ontology Consortium)创建的一套动态的、受控的等级结构词表,包括生物过程本体(Biological Process)、分子功能本体(Molecular Function)和细胞成分本体(Cellular Component)这 3 部相互独立的本体[22-23]。生物过程本体描述分子功能的有序组合,分子功能本体描述基因产物个体的功能,细胞成分本体描述亚细胞结构、位置和大分子复合物。GO 是一个有向无环图型(DAG)的本体,它有 3 个目标。一是保持并进一步发展基因及基因产物属性的受控词表,二是注释基因和基因产物,并吸收和传播注释数据,三是提供方便访问基因本体项目数据的工具[22]。基因注释工具有 GoGene(http://gopubmed.org/gogene)[24]等。支持在线或下载使用 GO 工程提供的数据的工具有基因本体联盟的 AmiGO[25]和 OBO - Edit[26-27],还有 GoPubMed[28], Comparative Toxicogenomics Database(CTD)[29]等,都很受欢迎。

2.2 命名实体识别

生物医学领域的命名实体包括基因名称、蛋白质名称、蛋白质结构属性名称、化合物名称、药物名称、疾病名称等。命名实体识别(Named Entity Recognition,NER)就是将其从文本数据中识别出来。其主要任务包括从文本中识别命名实体,确定该实体的类型,以及出现多个实体表示同一事物时,选择一个代表该组。它是信息抽取的基础,解决了这一问题将使更多复杂的文本挖掘问题迎刃而解[30]。

但是,生物医学命名实体识别比传统意义上的 NER 更具挑战性。例如,生物医学领域新名词的不断涌现,首字母缩写可以构成有效的基因名称,造成不同的基因具有相同的名称。获得性免疫缺陷综合征(Acquired Immune Deficiency Syndrome, AIDS)是一多词短语,现称作艾滋病等。这些问题使得 NER 变得复杂。

目前,解决方法有 3 种。一是基于字典的方法,就是与字典词条进行比对匹配,但是包含所有生物医学领域命名实体名称的词典是不存在的。二是基于规则的方法,就是按照定义的规则将实体与其他文本数据区分开来。三是基于统计的方法,就是从样例数据集合中统计出相关特征和参数,以此建立识别模型,最终识别出测验文本的命名实体,也称基于机器学习的方法。贝叶斯模型、隐马尔可夫模型(HMM)、支持向量机(SVM)、条件随机场(CRFs)、最大熵(Maximum entropy, ME)等[31]方法已经广泛用于识别有注释的语料集中的实体。当然,也可以综合以上方法进行命名实体的识别[32]。确定适当的特征模版和选择重要的特征值是这些方法成功的关键。杨志豪等[33]认为现在较为流行的机器学习方法仍需改善,并提出了一种基于条件随机域的生物医学 NER 方法,选取适当特征进行实体识别,利用上下文线索进一步提高识别性能。结果表明,上下文线索的引入,使识别性能在条件随机域方法基础上提高了近 3%。

可以进行命名实体识别的系统很多,较为著名的有 ABNER[34], BalIE[35]和 PASTA 等。SciMiner 是一个可进行 MEDLINE 文摘及全文分析的基于网络的文本挖掘及功能分析工具,接受自然语言提问和 PubMed 标识符输入。使用的方法是规则表达模式及字典方式,首字母缩写引起的歧义通过一个基于首字母共现及相应说明条款的评分方案解决。功能分析用于识别高度相关目标(基因和蛋白质)。GO 术语、MeSH 术语、通道和蛋白质相互作用网络,使用 BioCreative II 评价基因/蛋白质名称识别,得到 87.1% 的查全率、71.3% 查准率及 75.8% 的 F - 测度[36]。Sujan Kumar Saha 等[37]开发了一种基于最大熵的生物医学 NER 系统,识别和选择的功能主要是自动完成。与其他生物医学 NER 系统相比没有使用深层次的领域知识,使用最大熵分类器构建系统,用 JNLPBA 2004 训练语料集训练分类器,采用 JNLPBA 2004 测试数据评价系统,得到了较高的查准率、查全率及 F - 测度。采用 JNLPBA 2004 数据的训练和测试系统见表 1。

表 1　几种采用 JNLPBA 2004 数据的训练和测试系统[37]

系统	机器语言方法	领域知识	F - 测度
Sujan et al. (2008)	MaxEnt	POS information	67.41
Zhou & Su (2004) Final	HMM, SVM	Resolution of Name alias, Cascaded NEs, Abbreviations; Dictionary; POS	72.55
Zhou & Su (2004)	HMM, SVM	POS information	64.10
Song et al. (2004) Final	SVM, CRF	POS Information, Phrase, Virtual Sample	66.28
Song et al. (2004) Base	SVM	POS Information, Phrase	63.85
Ponomareva et al. (2007)	HMM	POS information	65.70

此处需要说明的是,评价时使用不同语料集,其评价结果是不能比较的。如 Shen D 等[38]开发的基于 HMM 的 NER 系统,在 GENIA V1.1 中是 62.5%,在 GENIA V3.0 中则是 66.1%。

2.3　关系抽取

关系抽取(Relationship extraction, RE)的目标是检测一对特定类型的实体之间有无预先假设的关系[8]。生物医学文本挖掘抽取的就是基因、蛋白质、药物、疾病、治疗之间的关系。如 HSP 蛋白主要和视网膜疾病有关,可能导致白内障或者致盲。TNFα 的表达水平变化主要是和 2 型糖尿病的致病有关,具有致炎性、引发细胞坏死等作用[39]。

有人将应用于此项工作的方法归纳为基于模版的方式(手动、自动)、基于统计的方式、基于自然语言处理的方式。基于模版的方式就是从已知且有兴趣关系的实体周围的文本中归纳出模式,再利用这个模式对测试语料集的文本进行模式匹配。基于统计的方法就是通过寻找经常一起出现而常多于随机出现的实体而识别出关系。基于自然语言的方法就是把自然语言分解为可从中提取出关系的结构[8]。这些方法也可混合使用。

药物基因组学研究人类基因与药物反应之间的关系,识别药物和分子实体,尤其是基因和基因变异之间的重要关联。Pharmspresso 是可在全文中识别重要药物基因组学事实的文本挖掘工具,可挖掘人类基因、多态性、药物和疾病以及它们之间的关系,且评价良好[40]。蛋白质相互作用(Protein - protein interaction, PPI)是关系抽取的一个重要方向。Makoto Miwa 等人使用基于核心(Kernel - based)的机器学习方法来自动抽取蛋白质相互作用关系,在几种句法分析器的基础上结合内核,以便检索到尽可能广泛且重要的信息,使用支持向量机(SVM)评价这种方法,并取得很好的结果[41]。目前,大多数工具均使用基于共现的方法和基于规则的方法。中科院发育生物学中心的李巍等人认为,混合方法(基于框架的方法)结合这两种方法能够更好地预测 PPIs,但是这些方法很少被知名的蛋白质关系数据库及基因本体数据库中的共现术语评估。他们开发了一种基于 WEB 的工具——PPI Finder (http://liweilab.genetics.ac.cn/tm/),可通过分析 PubMed 摘要中共现及一起活动的单词,挖掘出人类的 PPI。结果发现,PubMed 摘要中的共现词对只有 28% 出现在常用人类 PPI 数据库中(HPRD, BioGRID 和 BIND),已知的 HPRD 中的 PPIs,有 69% 在文献中共现,65% 共用基因本体术语[42]。

关系抽取系统的侧重点各有不同。如 GENIES (Genomics information extraction system)侧重于提取参与细胞途径的已知基因的关系,EDGAR 系统侧重药物与癌症相关基因的关系,PSI - BLAST(Position Specific Iterated Basic Local Alignment Search Tool)侧重于以文件比较算法检测已知结构的远程同源染色体[43]。

2.4　文本分类和聚类[2,44-45]

2.4.1　文本分类

文本分类(Text classification)就是将文本自动归入预先定义好的主题类别中,是有监督的机器学习方法,主要应用于自动索引、文本过滤、词义消歧(WSD)和 Web 文档分类等。一般来说文本分类需要获取训练文本集、选择分类方法并训练分类模型、用分类模型对其他文本进行分类和根据分类结果评估分类模型等 4 个步骤。

目前,文本分类的方法有很多,典型且效果较好的有朴素贝叶斯分类法(Na Bayes)、K 最近邻(K - NN)、支持向量机(SVM)、决策树等,还有基于关联

的分类(CBA)及基于关联规则的分类(ARC)。Eskin E [46] 使用 SVM 算法和基因序列 kernel 预测蛋白质在细胞质中的位置,达到了 87 % 的查准率和 71% 的查全率。Conway M 使用 n - grams 和语义特征对疾病爆发报道进行了文本分类: 特征集——命名实体特征、n - grams 特征和来自 USAS 语义标注器的特征,方法是 3 种机器学习方法——朴素贝叶斯、支持向量机算法、C4. 5 决策树算法,特征选择使用 chi(2)特征选择算法。研究表明,此方法从统计水平上提高了分类精度 [47]。Ambert KH 等 [48] 开发了一种从临床报告信息中有效分类合并症的文本挖掘系统,使用了多种自动化技术,包括热点过滤、否定概念识别、零矢量滤波,加权的逆级频率,以及纠错码输出与线性支持向量机。

2.4.2 文本聚类

文本聚类(Text clustering)是根据文本数据的特征将一组对象集合按照相似性归纳为不同类的过程,与文本分类的区别是分类的对象有类别标记。而聚类是根据聚类算法自动确定的,又被称为无监督的机器学习方法。文本聚类的步骤为获取结构化的文本集,执行聚类算法,获取聚类谱系图,选取合适的聚类阈值。

常见的聚类算法可归纳为平面划分法(如 K - 均值算法、K - 中心点算法),层次聚类法(可分为凝聚层次聚类和分割聚类),基于密度的方法(如 DBSCAN 算法),基于网格的方法(如 STING 算法),基于模型的方法。以上方法可以综合使用,如 CLIQUE 就是综合了密度和网格两种聚类方法。生成高质量的基因簇和识别基因簇的基本生物学机制是聚类基因表达分析的重要目标。Hu X [49] 设计并开发的 GE - Miner 通过文本聚类和多文档总结,提供一个综合分析基因表达数据的分析环境,使用 GE - Miner 可获得高质量的簇和基因簇的文字简介。PuReD - MCL(PubMed Related Documents - MCL)是一个基于图的 PubMed 的文档聚类方法。它不直接使用自然语言处理技术,而是充分利用 PubMed 现有资源,然后使用基于流动模拟图的 MCL 图聚类算法。这一过程允许用户利用重要线索分析结果,最终使用交互式图形布局算法,可视化群簇和所有相关信息 [50]。Groth P 等 [51] 根据显型的描述,利用文本聚类将基因聚类成簇,利用这些簇预测基因功能,采用客观标准选择一个子类团,从生物过程次本体中预测 GO - 术语注释,得到了 72. 6% 的查准率和 16. 7% 的查全率。国内也有很多根据聚类方法进行生物医学文本挖掘方面的研究,如以急性白血病和急性髓样白血病为研究主题,通过对主题词共现关系进行聚类,挖掘相关文献中的基因与疾病之间的关系等 [52]。

2.5 共现分析 [5,54]

共现(Co - occurrence)分析主要是对隐性知识的挖掘,在生物医学领域主要用于诸如 DNA 序列的数据分析、基因功能相似聚类、基因和蛋白质的功能信息提取、提高远程同源性搜索、基因与确定疾病关系预测等 [6]。自然语言的模棱两可及其极大的灵活性,导致自然语言处理系统把重点放在了生物医学领域非常具体的问题上,比如蛋白质间的相互作用。共现分析则截然不同,它通过关注单词和术语的分布,或者通过分析不同段落的信息内容,在现有知识中发现“出人意料”的联系或问题,最终可能会产生新的研究方法或知识。具体来说,如果在大规模语料(训练语料)中,两个词经常共同出现(共现)在同一窗口单元(如一定词语间隔、一句话、一篇文档等)中,则认为这两个词在语义上是相互关联的。而且,共现的频率越高,其相互间的关联越紧密。基于这样的一个假定,通过对训练语料的统计,计算得到词与词之间的互信息(Mutual information),就可以对词与词之间的相关性进行量化比较,获得对文本词汇语义级别的关联认识。王曰芬等 [53] 的共现分析文本挖掘流程如图 2 所示。

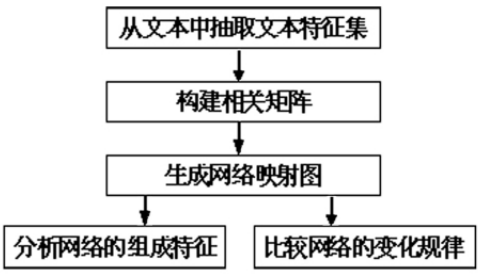

图 2 共现分析文本挖掘流程

构造相关矩阵和建立词汇向量主要涉及关联度的计算。关联度是对以分散属性为特征的对象之间相关性的测度,计算词汇间的关联度可以分析词汇之间的关系,也可以将词汇聚类构成概念,计算概念之间的关联度。主要方法有 Dice 指数、余弦指数 [54]。由于 Jaccard 指数能够根据词的共现频率直接反映两个词之间的相似度并且消除高频词的消极

影响，它被广泛用作两词之间的标准化相关系数。对利用上述方法计算出的关联度，聚类词汇成簇，其中每个簇代表一个研究子领域。要确定领域内部关联关系和核心领域，并比较不同时期的簇（图），发现研究领域的发展变化，主要评价指标有包容指数（Inchistion index）、邻近指数（Proximity index）和等价指数（Equivalence index）3 种。包容指数主要用来突出领域中的中心概念并对其关系进行描述，邻近指数用以确定发展迅速的次要研究领域，等价指数是用以计算词汇相互包含关系的指数。生成网络映射图是为帮助人们更直观的理解，即可视化问题。映射图生成过程的关键在于统计分析方法（如系统聚类、主成分分析和多维尺度）的选择。当然，可以同时采用多种统计分析方法揭示数据内涵。Müller H 等人使用 NetCutter 识别和分析共现网络并且发现一个新蛋白质与一组基因有关，这组基因经常协调限制人类癌症相关基因的表达。NetCutter 可应用于任何一套可通过共现分析揭示功能关系的数据，如临床参数与癌症亚型的联系，或单核苷酸多态性与疾病表型的联系[55]。最新的文本挖掘工具 PPI finder[56]，SciMiner[56]，BioCreative II[57] 也都用到了共现分析技术。

2.6　系统（工具）、评价及结果可视化[58]

2.6.1　生物医学文本挖掘工具

完美的文本挖掘系统应该以生成“某种形式的命题”作为其输出结果，但是目前没有一个系统有如此先进的推理能力来完成这样的任务。问答系统是类似这样一个准确率较高的系统，即以自然语言方式提问，采用自然语言处理技术，自动返回给用户简短、具体的答案，但仍处于设想阶段[4]。

目前已有多种免费的文本挖掘系统，有学者已对它们的功能进行了简单比较（表 2）。综合比较各挖掘系统可以看出，挖掘系统一般有文献检索模块、文本转化和结构化模块、自然语言处理模块、文本挖掘模块 4 个功能模块。这 4 个模块需要完成生物医学文本挖掘的各项任务。

表 2　常见工具比较[58]

工具	PDF	基本处理	词典	规则	机器学习	关系抽取	语料库导航
ABNER			√			√	
AliBaba		√	√			√	√
BiolE			√		√		
Chilibot				√		√	√
EBIMed		√	√			√	√
EDGAR			√	√		√	
EMPathIE		√	√	√		√	
GAPSCORE		√		√	√		
GeneWays		√	√	√		√	√
GIS		√	√			√	
GoPubMed		√	√				√
iHOP		√	√				√
LitMiner			√				√
MedEv			√	√			√
MedGene			√			√	√
MedIE			√		√	√	
MedMiner		√	√			√	√
PaperBrowser	√	√	√		√		
PASTA		√	√	√			
PolySearch		√	√	√			√
POSBIOTM/W				√	√	√	
PubGene			√			√	
PubMatrix			√				
QUOSA							
Suiseki		√	√	√		√	
Textpresso	√			√		√	
TIMS			√	√			

注：基本处理是基本的预处理过程，包括 tokenisation、stemming、停用词消除、句子划界等。基于词典、基于规则、基于机器学习方法都是用以命名实体识别。

2.6.2 评价

文本挖掘系统建立以后，需要对其进行评价。系统开发阶段通常使用领域专家手工标注过的标准语料库（金标准），检验和测试系统的性能指标，较为流行的金标准语料库是 GENIA 语料库[59]。文本处理结果准确度常用查全率和查准率两个指标衡量。查全率（Recall）= 系统输出的结果与实体相关的结果数/标准语料库中与实体相关的结果数，即“我找到多少正确的结果”。例如，想要寻找谈及 A 的文章，查全率就是在所有涉及 A 的文章中，检索到的最大比例。查准率（Precision）= 系统输出的结果中与实体相关的结果/系统输出的结果总数，即“我认为是正确的结果有多少是真正正确的”，例如，在上述所涉及 A 的文章中，有多少是确实满足检索者期待而不是被错误划分进来的。

系统的评价还应关注系统的稳定性问题。如果说准确的处理少数文本文件不算很难，那么每天处理几百万篇文章，对整个软件系统来说就是一大考验了。算法的高效、系统的稳定、可扩展性都起着决定性作用[60]。

2.6.3 可视化

可视化（Visualization）技术通常不是文本挖掘的一部分，但是由于用户必须处理大量的文本数据，并且常常出现复杂的浏览结果，大部分系统不得不提供新的方法来扩大可视化结果，使用户的挖掘更具效率。如 GeneWays 系统可显示分子网络，SemGen 系统、g2p 系统和 PGviewer 可显示基因与表型的关系等。可视化文本挖掘就是把可视化技术引入文本挖掘过程，充分利用人们对可视模块快速识别的自然能力，通过有效的可视界面，让人们能够理解大规模的数据。信息可视化的实际应用包括选择、转换、表现抽象数据在某一形式中，这种形式可以使查找和理解人机交互。视图的交互和动态是信息可视化的重要方面。

简单地说，可视化模块就是用可视化算法将文本挖掘结果用图形表现出来。图中节点是一个关键信息的代表，连线表示它联结的两个节点间的关联，连线的粗细表示节点间的关联度，颜色表示两节点同时出现的频率。当从一个关键信息关联到另外一个关键信息时，系统就绘制一个新的节点，并用一条连线连接开始节点和新的节点[61]。AKS（Almaknowledge Server）系统给出的共现分析图（图 3），显示了阿尔茨海默病（老年性痴呆）和 ACHE，APOE，APP 基因的直接关联，也列出了疾病有关的化合物。一组与 ACHE 有关的化学物质是这种酶的抑制剂，适合治疗阿尔茨海默病。多奈哌齐和他克林也与 APOE 有关联。这种基因的不同基因型可能影响药物活性。修饰机制显示与 APP 基因有关[61]。

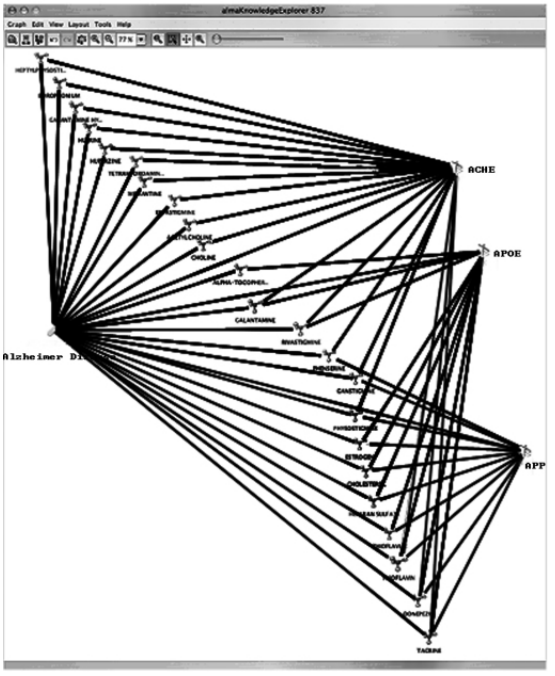

图 3 共现分析图

3 结语

目前国内在生物医学文本挖掘方面的研究不是很多。文本挖掘方面有一个较热的关注点就是中文分词问题，难点是汉字字符与西方语言不同，是没有边界的。全文挖掘也是一个亟待解决的问题，主要因为商业原因对文本全文的限制性访问，所以大部分生物医学文本挖掘都是从数据库的摘要、标题、关键词等入手。

【参考文献】

[1] 杨建武. 文本挖掘技术：文本挖掘工具与应用 [EB/OL]. [2009-08-20]. http://cn.minidx.com/index.php? option=com_docman&task=doc_view&gid=44.
[2] 王丽坤，王 宏，陆 玉. 文本挖掘及其关键技术与方法 [J]. 计算机科学，2002，29（12）：12-19.
[3] Cohen AM, Hersh WR. Hersh. A Survey of Current Work in Biomedical Text Mining [J]. Brief Bioinform（S1467-5463），2005，6（1）：57-71.
[4] 张 智，张正国，蛋白质相互作用的文本挖掘研究进展 [J]. 中国生物医学工程学报，2008，27（5）：764-782.
[5] 齐 彬，吕 婷. 共现分析技术在生物医学信息文本数据挖掘中的应用 [J]. 中华医学图书情报杂志，2009，18（3）：41-43.
[6] Erhardt RA, Schneider R, Blaschke C. Status of text-mining tech-

niques applied to biomedical text [J]. Drug Discov Today(S 1359 - 6446),2006,11(7/8):315 - 325.
[7] Denny JC, Miller RA, Waitman LR, et al. Identifying QT prolongation from ECG impressions using a general - purpose Natural Language Processor [J]. Int J Med Inform(S1386 - 5056),2009,78(S1):34 - 42.
[8] Novichkova S, Egorov S, Daraselia N. MedScan, a natural language processing engine for MEDLINE abstracts [J]. Bioinformatics(S1367 - 4803),2003,19(13):1699 - 1706.
[9] Rzhetsky A, Iossifov I, Koike T, et al. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data [J]. J Biomed Inform(S1532 - 0464),2004,37(1):43 - 53.
[10] Spasic I, Ananiadou S, McNaught J, et al. Text mining and ontologies in biomedicine: making sense of raw text [J]. Brief Bioinform(S1467 - 5463),2005,6(3):239 - 251.
[11] Chen L, Liu H, Friedman C. Gene name ambiguity of eukaryotic nomenclatures [J]. Bioinformatics(S1367 - 4803), 2005, 21(2): 248 - 256.
[12] Friedman C, Shagina L, Lussier Y, et al. Automated encoding of clinical documents based on Natural Language Processing [J]. J Am Med Inform Assoc(S1067 - 5027),2004,11(5):392 - 402.
[13] Tulipano PK, Tao Y, Millar WS, et al. Natural language processing and visualization in the molecular imaging domain [J]. J Biomed Inform(S1532 - 0464),2007,40(3):270 - 281.
[14] 袁　毅，张　丹，张晓东，等. 基因相关生物医学文献挖掘研究 [J]. 电脑知识与技术，2008(13):620 - 623,677.
[15] Bodenreider O. The Unified Medical Language System(UMLS): integrating biomedical terminology [J]. Nucleic Acids Res(S0305 - 1048),2004,32(Database issue), D267 - D270.
[16] UMLS. [EB/OL]. [2009 - 08 - 20]. http://www.nlm.nih.gov/research/umls/.
[17] MetamorphoSys. [EB/OL]. [2009 - 08 - 20]. http://metamorphosys.org/.
[18] MetamorphoSys - The UMLS Installation and Customization Program [EB/OL]. [2009 - 08 - 20]. http://www.nlm.nih.gov/research/umls/meta6.html.
[19] Lexical Variant Generation(LVG) [EB/OL]. [2009 - 08 - 20]. [2009 - 08 - 20]. http://www.nlm.nih.gov/research/umls/online_learning/LEX_004.htm [EB/OL].
[20] Aronson AR. Comparison of LVG and MetaMap Functionality [EB/OL]. [2009 - 08 - 20]. http://skr.nlm.nih.gov/papers/references/LVG - MetaMap.comparison.pdf.
[21] The Gene Ontology project. [EB/OL]. [2009 - 08 - 20]. http://www.geneontology.org/.
[22] Gene Ontology Consortium. The Gene Ontology project in 2008 [J]. Nucleic Acids Res(S0305 - 1048), 2008, 36(Database issue):D440 - 444.
[23] The Sequence Ontology Project. EB/OL]. [2009 - 08 - 20]. http://www.sequenceontology.org/.
[24] Plake C, Royer L, Winnenburg R, et al. GoGene: gene annotation in the fast lane [J]. Nucleic Acids Res(S0305 - 1048),2009,37(Web Server issue):W300 - 304.
[25] Gene Ontology. AmiGo [DB/OL]. [2009 - 08 - 20]. http://amigo.geneontology.org/cgi - bin/amigo/go.cgi.
[26] SourceForge [EB/OL]. [2009 - 08 - 20]. http://sourceforge.net/project/showfiles.php? group_id = 36855.
[27] Day - Richter J, Harris MA, Haendel M, et al. OBO - Edit - - an ontology editor for biologists. [J]. Bioinformatics(S1367 - 4803),2005,23(16): 2189 - 2000.
[28] GoPubMed. [DB/OL]. [2009 - 08 - 20]. http://www.gopubmed.org.
[29] The Comparative Toxicogenomics Database (CTD) [DB/OL]. [2009 - 08 - 20]. http://ctd.mdibl.org/.
[30] The GO Consortium (2009 - 03 - 16). "gene_ontology.1_2.obo" (OBO 1.2 flat file) [EB/OL]. [2009 - 08 - 20].. http://www.geneontology.org/ontology/obo_format_1_2/gene_ontology.1_2.obo.
[30] Hanisch D, Fluck J, Mevissen HT, et al. Playing biology's name game: identifying protein names in scientific text [EB/OL]. [2009 - 08 - 20]. http://helix - web.stanford.edu/psb03/hanisch.pdf.
[31] Polajnar T. Survey of Text Mining of Biomedical Corpora [EB/OL]. [2009 - 08 - 20]. http://www.brc.dcs.gla.ac.uk/tamara/surveyoftm.pdf.
[32] Fundel K, Güttler D, Zimmer R, et al. A simple approach for protein name identification: prospects and limits [J/OL]. BMC Bioinformatics(S1471 - 2105),2005,6(S1):15. [2009 - 08 - 20]. http://www.biomedcentral.com/content/pdf/1471 - 2105 - 6 - S1 - S15.pdf.
[33] Yang Z, Lin H, Li Y. Exploiting the contextual cues for bio - entity name recognition in biomedical literature [J]. J Biomed Inform(S1532 - 0464),2008,41(4):580 - 587.
[34] Tsuruoka Y, Tateishi Y, Kim JD, et al. Developing a robust part - of - speech tagger for biomedical text [C]. Adv Inform Proc, Berlin / Heidelberg: Springer, 2005(3746):382 - 392.
[35] Smith L, Rindflesch T, Wilbur WJ. MedPost: a part - of - speech tagger for bioMedical text [J]. Bioinformatics (S1367 - 4803), 2004,20(14):2320 - 2321.
[36] Hur J, Schuyler AD, States DJ. SciMiner: web - based literature mining tool for target identification and functional enrichment analysis [J]. Bioinformatics(S1367 - 4803),2009,25(6):838 - 840.
[37] Saha SK, Sarkar S, Mitra P. Feature selection techniques for maximum entropy based biomedical named entity recognition [J]. J Biomed Inform(S1532 - 0464) (Epub ahead of print).
[38] Shen D, Zhang J, Zhou GD, et al. Effective adaptation of a hidden markov model - based named entity recognizer for biomedical domain [EB/OL]. [2009 - 08 - 20]. http://www.comp.nus.edu.sg/~tancl/Papers/ACL03/shen03acl.pdf.
[39] 徐　昊，陶　林，魏　武，等. 文本挖掘技术在整合蛋白与疾病关系资源中的应用 [J]. 生物信息学，2009,7(1):21 - 24.
[40] Garten Y, Altman RB. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text [J/OL]. BMC Bioinformatics(S1471 - 2105),2009,10(S2): 6. [2009 - 08 - 20]. http://www.biomedcentral.com/content/pdf/1471 - 2105 - 10 - S2 - S6.pdf.
[41] Miwa M, Soetre R, Miyao Y, et al. Protein - protein interaction extraction by leveraging multiple kernels and parsers [J]. Int J Med Inform(S1386 - 5056) (Epub ahead of print).
[42] He M, Wang Y, Li W. PPI finder: a mining tool for human protein - protein interactions [J]. PLoS One(S1932 - 6203),2009,4(2):e4554.
[43] Tulipano PK, Tao Y, Millar WS, et al. Natural language processing and visualization in the molecular imaging domain [J]. J Biomed Inform(S1532 - 0464),2007,40(3):270 - 281.
[44] 夏咏梅. 基于文本挖掘的分类与聚类技术 [J]. 情报探索，2005(3):65 - 67.
[45] 陈晓云. 文本挖掘若干关键技术研究 [D]，上海：复旦大学，2005.
[46] Eskin E, Agichtein E. Combining text mining and sequence analysis to discover protein functional regions [C]. Altman RB, Dunker AK, Hunter L, et al. Pac Symp Biocomput, 2004:288 - 299.
[47] Conway M, Doan S, Kawazoe A, et al. Classifying disease outbreak reports using n - grams and semantic features [J]. Int J Med Inform(S1386 - 5056), (Epub ahead of print).
[48] Ambert KH, Cohen AM. A System for Classifying Disease Co - morbidity Status from Medical Discharge Summaries Using Automated Hotspot and Negated Concept Detection [J]. J Am Med Inform Assoc(S1067 - 5027), (Epub ahead of print).
[49] Hu X. GE - Miner: integration of cluster ensemble and text mining for comprehensive gene expression analysis [J]. Int J Bioinform Res Appl(S1744 - 5485),2006,2(3):325 - 338.
[50] Theodosiou T, Darzentas N, Angelis L, et al. PuReD - MCL: a graph - based PubMed document clustering methodology [J]. Bioinformatics(S1367 - 4803),2008,24(17):1935 - 1941.
[51] Groth P, Weiss B, Pohlenz HD, et al. Mining phenotypes for gene function prediction [J/OL]. BMC Bioinformatics(S 1471 - 2105), 2008, 9: 136. [2009 - 08 - 20]. http://www.biomedcentral.com/content/pdf/1471 - 2105 - 9 - 136.pdf.
[52] 闫　雷，崔　雷. 急性白血病相关基因的文本挖掘分析 [J]. 情报学报，2008,27(2):169 - 174.
[53] 王曰芬，宋　爽，熊铭辉. 基于共现分析的文本知识挖掘方法

研究 [J]. 图书情报工作,2007,57(2):66 - 79.
[54] Ding Y, Chowdhury GG., Foo S. Bibliometric cartography of information retrieval research by using co - word analysis [J]. Inf Process Manag(S0306 - 4573),2001,37(6):817 - 842.
[55] Müller H, Mancuso F. Identification and analysis of co - occurrence networks with NetCutter [J]. PLoS One(S1932 - 6203),2008,3(9):e3178.
[56] He M, Wang Y, Li W. PPI finder: a mining tool for human protein - protein interactions [J]. PLoS One(S1932 - 6203),2009,4(2):e4554.
[57] Krallinger M, Leitner F, Rodriguez - Penagos C, et al. Overview of the protein - protein interaction annotation extraction task of BioCreative II [J]. Genome Biol(S1465 - 6906),2008,9(S2):4.
[58] Lourenco A, Carreira R, Carneiro S, et al. @ Note: A workbench for Biomedical Text Mining [J]. J Biomed Inform(S1532 - 0464),2009,42(4):710 - 20.
[59] Kim JD, Ohta T, Tateisi Y, et al. GENIA corpus—a semantically annotated corpus for bio - textmining [J]. Bioinformatics(S1367 - 4803),2003,19(S1):i180 - 182.
[60] 文本挖掘技术在 CIC 的应用 [EB/OL]. [2009 - 08 - 20]. http://blog.csdn.net/CICTech/archive/2008/04/16/2296453.aspx.
[61] 田栓成,刘金媛. 可视化竞争情报的提取 [J]. 情报杂志,2007,26(5):16 - 18.

[收稿日期:2009 - 08 - 05]
[本文编辑:王 颖]

(上接第 23 页)

事实上,如果一个人会使用图书馆的话,利用谷歌查找信息也会很有帮助,起码会帮检索者提高效率。由于谷歌和图书馆能够满足的范围不一样,谷歌可以提供的,图书馆一样可以提供;而谷歌不能提供的,图书馆也可以提供,所以,谷歌不会取代图书馆[10]。而网络作家 Will Sherman 不认为图书馆是"夕阳产业",并且"网络不能找到所有东西,很难分辨资讯的精华。网络是图书馆的补充,不会取代图书馆。图书馆是稳定存在的,网络资讯则可能稍然即逝"[9]。

谷歌作为一种检索工具,其作用是有限的。例如,用它漫无边际地检索出一大堆无用的噪声信息是不能令人满意的。因此,在这种情况下专业数据库或专业检索工具的优势就明显地显现来了。此外,一些未对谷歌开放的网站与数据库是对谷歌检索功能权威性的最大挑战。信息量大并不能等同于信息全,这也是影响谷歌检索质量的一个重要因素。

目前,谷歌公司的图书数字化项目进行得并不顺利,其最大原因就是版权因素。因此,很多业界人士就版权问题对谷歌进行了严厉批判,导致谷歌的计划受挫,甚至上述五大图书馆也只能向谷歌提供版权保护期届满的图书。

4 结语

谷歌是网络时代的必然产物[11]。谷歌为图书馆界带来的不仅仅是冲击,也有很好的促进作用。例如,从技术层面上来看,它提供了方便快捷的搜索途径,解决了长久以来一直困扰人们的许多技术难题。它的索引、检索结果排序、书刊数字化等方面的新技术,引领了整个信息时代,它在自然语言方面也无可匹敌。该网站推出的许多产品都是人们很早之前就苦思寻觅而求之不得的[12]。这正是图书馆早应该考虑的问题。

作为具有专业知识背景和提供信息服务的图书馆员,应该充分利用谷歌这一方便快捷的搜索工具,有效地评估信息质量,向读者或用户提供更权威、有深度的知识和信息。

谷歌与百度等网站的出现,对于图书馆界来说既是机遇又是挑战。图书馆要不失时机地抓住机遇,调整心态,适应转变,找准定位,练好内功,着力发展自身的长项和强项。面对新的挑战,图书馆界要充分利用谷歌和百度等学术网站并开发和挖掘其服务功能,积极、主动向用户提供个性化、有针对性的深层次的信息服务,在与谷歌、百度等网络公司的竞争中共存和发展。

【参考文献】

[1] 康 娟. 谷歌首推虚拟图书馆 可能引发"革命" [N]. 解放日报,2005 - 11 - 05.
[2] 陆 鹏. 五个层面分析 Google 兵败中国图书馆项目 [EB/OL]. [2006 - 07 - 17]. http://home.donews.com/donews/article/9/98724.html.
[3] 编目精灵. 取代 OPAC:今天是超星、明天是谷歌图书 [EB/OL]. [2008 - 10 - 31]. http://catwizard.blogbus.com/logs/30851899.html.
[4] 邓佩珍. 在竞争中谋求共存和发展:谷歌数字图书馆计划对图书馆的挑战 [J]. 晋图学刊,2008(4)32 - 35.
[5] Keven. 关于 2.0 迷惑 [EB/OL]. [2008 - 05 - 25]. http://www.kevenlw.name/archives/category/%E5%9B%BE%E4%B9%A6%E9%A6%86%8620.
[6] 胡小菁. 谷歌图书馆合作计划的背景、目的与分析 [J]. 图书馆杂志,2005(5):17 - 20.
[7] 张艳霞. 谷歌:图书馆网络服务新模式 [J]. 图书馆杂志,2008(8):47 - 49.
[8] 石剑峰. 图书馆不会被网络所取代 [EB/OL]. (2007 - 09 - 29) [2008 - 11 - 26]. http://sdqnb.dzwww.com/qingnianbaoshiban/200709/t20070929_2498832.htm.
[9] 灯火阑珊处. 数位化,图书馆还存在吗? [EB/OL]. (2009 - 04 - 06) [2009 - 06 - 08] http://blog.ifeng.com/article/2532980.html.
[10] 赵 斌. 谷歌,不是图书馆的终结者 [N]. 成都日报,2008 - 11 - 03.
[11] Hwlibrary 的 BLOG. 谷歌与图书馆 [EB/OL]. (2007 - 00 - 23) [2008 - 09 - 25]. http://blog.sina.com.cn/s/blog465ce7d901000a31.html. 2007 - 06 - 23.
[12] 翟晓娟,许 鑫. 浅论谷歌向学术领域的发展及其对图书馆的影响 [J]. 图书与情报,2008(3):61 - 66.

[收稿日期:2009 - 07 - 23]
[本文编辑:杜海洲]