自然科学文章信息检索

**（特别提醒：此题相关附件见人人网“南开数模协会”主页）**

**背景知识**

对于研究者来说，研究任何问题之前，都要对当前的研究状况作充分的了解，以便确定要研究的问题是否有新意？是否可行？找到合适的文献是一个基本的但也是本质的工作，这就是所谓的信息检索（information retrieving）。当然这是一个耗时的工作。到目前为止，这项工作还是研究者本人或者助手人工进行来完成的。如果交给一个助手来完成，那么由于助手的知识和眼光的局限很可能得不到全面的信息。如我们所知，当前的google、百度以及其他搜索引擎搜索能力还是很不错的。但是，如果键入某些关键词之后，google搜索引擎会在google的文本数据库中迅速找出很多文献，而研究者必须根据每篇文章的内容才能确定是否为需要的，阅读量是在太大了。

在当今的大数据时代，信息检索已经是计算机科学不能绕开的问题之一。有人依据每篇文章的关键词作为信息源，设计了基于关键词的搜索工具来帮助研究者监控他们的领域内网页或者文献库的变化。随着云计算技术的推进，云搜索也就变得很重要了，它可以保证研究者获得充分的文本来构建自己的文献数据库。不会因为局限在某个不全的文献库中找不到就认为不存在。但是由于作者在发表论文时对于关键词提供没有具体要求（只有个数限制），因此从关键词到最终知识获取还有很大距离。需要将依据关键词搜索到的全部文献中的与所要的知识无关的内容过滤，只保留研究者所需的文献。

在医学领域，特别是针对一些疾病，不论是常见病还是罕见病，都会在文献检索方面遇到难题。比如，对于常见病（癌症、高血压、心衰、脓毒症），文献虽多，但所需的文献可能还不在常见的某个文献库中。对于罕见病，本来可能研究文献就少且不一定在流行的文献库中，甚至还有可能使用了别名，于是就出现了“查不到”的尴尬局面。因此，往往需要根据相同的疾病症状或者其他特征来查询。特别当这个参考的疾病的起因、或者相关的基因与机制都已经知道的话，对于该罕见疾病的研究会具有相当大的帮助。这就需要用户持续更新文献数据库，使得最新、最重要的深入结果不断加入到只有该领域专家才懂的文献库中。

# 为了克服因为没有使用学名或者查询方式不妥而导致“查不到”，需要建立一个纲目（Ontology），使其包含疾病学名和所有别名。这个纲目应该包括较为详细的内容比如，疾病学名（别名）、临床症状、起病原因、生物标记物、酶、基因、病理机制、治疗药物等，但一般是逐渐增补的，除非是研究透彻的疾病，一般不太可能一蹴而就。例如，当仅知道一个疾病的学名，其他什么也不知道时，先从wiki上得到该学名的所有别名，然后从文献库中搜索所有包含学名或者某种别名的所有文献，然后从这些文献中获取所有涉及到的临床症状，通过临床研究的经验确认哪些症状是最可能的症状加入到Ontology中。再基于新的Ontology再去检索与你检索的疾病具有相同症状的疾病，从其中研究透彻的某种疾病中发现起病原因、生物标记物、分泌物、酶、分子机制、治疗药物逐步添入Ontology。Ontology可以由用户或者临床大夫自己提供，也可以当用户给出一点线索之后，编写逐步更新程序来实现。

# 但是筛选哪些是哪些不是用户真正想所要的，就不那么简单了。现有的一些方案，比如Textpresso，这是一个基于纲目的针对生物文献的信息检索和提取系统。效果也虽然不是很理想，但目前还没有更好的方法发表。原因是它仅依赖于较少的关键词，这样的信息源太窄。由于文章的关键词与实际内容关系不是很大。因为每个作者发表文章时并不十分在意关键词的推敲，其实除了数量有一定限制，没有硬性规定，因此可能不同文章会选择相同的关键词。相比之下，每篇文献的内容摘要还是很能够恰当地反映文章的全貌。因为审稿时有一项就是判断摘要是否写得合理。不同文章不可能对应相同的摘要。特别是生命科学类的文章，摘要一般包含背景、方法、结论。于是，将摘要作为信息源，就能很好的揭示该文献。如果将摘要去掉标点符号，改成fast格式，那么摘要就是一个如同氨基酸序列的文本文件。因为20个标准的氨基酸的单字母表示分别为A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y. 对于多出的六个字符B, J，O, U, X，Z认为是可以忽略的符号，那么也可以采用蛋白质序列比对的算法mcabmsa.20130520来进行快速比对。为此，需要对于生物医学文章中的词汇的长度分布，平均长度进行统计，以便获取使用mcabmsa的参数。

# 任务：针对具体的疾病，Alzheimer Disease，做如下的信息检索（retrieving）工作:

# 采用当前流行某种云搜索引擎，建立动脉硬化的完备的文献库（只包含文献来源的网页、文件名、摘要）。并进行如下预处理：

# 统计文献库中单词的长度的分布以及平均长度。

# 给出将文献的title以及摘要改为适合于mcabmsa 处理的fast格式的程序。即

# >网页名缩写（空格）文献标题（空格）、发表日期、杂志名称 作为一行。

# 从abstract后的第一个字符开始到key words 之前结束的段落去除标点符号以及空格作为一行。

# 2）以平均长度L作为mcabmsa程序中锚点长度的参数，松弛选择为k=L\*10/13, 将fast 格式的文本库进行比对，输出是若干个集合，称为一个类，每类中包含至少一条信息。检验每类中是否反映相同或者相似的结果？

# 3）根据这些输出集合建立最优Ontology。