基于领域本体的生物医学语义检索机制研究
——以GoPubMed和SEGoPubMed为例

□ 杜建 张士靖* / 华中科技大学同济医学院医药信息系 武汉 430030

摘要：基于领域本体对大型文本语料库的语义检索研究成为近年来信息检索领域研究的热点。GoPubMed和SEGoPubMed作为该领域研究成果的杰出代表，为PubMed存在的问题提供了良好的解决方案。对GoPubMed和SEGoPubMed语义检索机制进行深入研究后显示，GoPubMed应用基于语义网的语义分类工具——GO和MeSH，将PubMed检索结果统计并对应到相应的分类导航目录，使用户利用该导航便可以快速找到自己所需要的文献，但其仍是基于布尔逻辑模型的字符串匹配机制；而比GoPubMed性能更优越的SEGoPubMed应用潜在语义分析（LSA）技术开创了语义敏感地浏览、连接大型文本语料库与本体的新模式，将具有更广泛的应用前景。

关键词：领域本体，基因本体，医学主题词表，GoPubMed，SEGoPubMed，潜在语义分析
DOI：10.3772/j.issn.1673—2286.2010.07.012

1 引言

21世纪，生命科学与医学成为科技发展的热点与前沿，MEDLINE/PubMed作为目前世界公认的最权威的生物医学文献数据库，收录的文献量已近2,000万条[1]。尽管其自动词语匹配功能令人称道，数据挖掘和智能检索功能日臻完善[2]，但从如此庞大的数据库中检索到所需文献仍然是一个巨大的挑战。因为PubMed以布尔逻辑为基础、基于关键词的检索模型以及检索结果输出的线性排序机制有两个缺点：（1）要想选择能够产生较好的检索结果的关键词，用户需要非常熟知检索策略及其优化方法；（2）检索结果的线性排序导致了用户不能从多个维度浏览检索结果，而只能逐篇阅读筛选，这样往往容易产生大量的信息冗余和信息丢失。

因此，为解决上述问题，基于领域本体对大型文本语料库实现语义检索成为近年来信息检索领域的研究热点[3]。领域本体是一个对既定领域的知识进行结构化表示的受控词汇表，用以减少和消除概念及术语的混乱。它提供了对该领域知识的共同理解，并从不同层次的形式化模式上给出这些术语及术语间相互关系的明确定义[4]。GoPubMed和SEGoPubMed的出现，就是该领域研究成果的杰出代表，它为PubMed存在的问题提供了良好的解决方案。本文以此为例，探讨基于领域本体的生物医学语义检索机制，其目的就是阐述如何实现PubMed的语义检索，进一步提高其检索效率，为用户提供高信度、低冗余的文献集合。

2 GoPubMed语义检索机制

2.1 语义检索原理

GoPubMed（http://www.gopubmed.org/）是德国Transinsight公司与德累斯顿工业大学（Technical University Dresden）合作研发的一种基于领域本体、对PubMed进行文本挖掘、结果可视化和后处理类型的生物医学语义搜索引擎[5]。GoPubMed以PubMed作为其来源数据库。在用户界面输入检索词后，系统将该词提交给PubMed，再接收从PubMed返回的检索结果（目标文本）；利用GO（Gene Ontology，基因本体）、MeSH（Medical Subject Headings，医学主题词表）以及UniProt（Universal Protein Resource，统一蛋白质资源）进行本体术语与目标文本的映射（术语提取）；通过文本挖掘，与查询请求相关的所有本体术语（GO

* 张士靖为通讯作者。

术语、MeSH词、UniProt蛋白质名称）均被提取出来，形成诱导本体（Induced Ontology），且诱导本体中术语之间的结构与基因本体以及MesH词表中术语之间的既有结构保持一致。由于每个GO术语都已经进行了注释，而且形成了一个术语（概念）、同义词、相关概念相互关联的初级语义网，系统将GO术语每个词汇的信息内容都考虑在内，这样在GO术语和摘要内容之间进行局部字序列（字符串）匹配时，如果在语义上一致，该术语即被自动提取出来。检索结果被统计并对应到由本体术语组成的相应的分类导航目录，用户利用该导航便可以快速找到自己所需要的文献。

2.2 领域本体

当前应用最广泛的生物本体是由基因本体协会（Gene Ontology Consortium）开发和维护的GO[6]。GO是一个应用GO术语（GO terms）对不同数据库中的基因和基因产物进行统一描述的受控词汇表。截至2009年4月6日，GO共收录了27,083条术语，对约20个生物数据库中超过1,600万条的基因和基因产物进行了注释[7]，且每天都在更新。基因本体协会将这些术语分为3个大类：molecular function（分子功能），biological process（生物过程）及cellular component（细胞成分），每一大类均已形成各自独立的子本体。每类按等级结构排列，层层向下的树型分支结构形成了一个有向无环图（directed acyclic graph，DAG），揭示了各物种的基因功能分类体系。

MeSH本体作为UMLS的典范，是美国国立医学图书馆（NLM）研制、逐年更新的受控医学主题词表。2009版MeSH收录本体术语（医学主题词）25,186个[8]，并分别按照字顺和树状结构排列。该表将所有主题词划分为16个大类，再逐级展开，最多可达11级。MEDLINE数据库应用MeSH本体进行了规范的主题标引，其自动词语匹配功能和数据挖掘技术的应用允许用户进行主题概念的扩展检索以及主题词/副主题词的聚焦检索。因此GoPubMed在接收PubMed检索结果时能够按照MeSH树形表对其进行自动分类，形成与检索提问相关的诱导本体（临时MeSH）。

UniProt是2002年由SIB、EBI、PIR基于信息整合和标准统一的需要，将Swiss-Prot、TrEMBL和PIR蛋白质序列数据库的三大巨头联合组建的一个新型蛋白质数据库[9]。这是当前最全面的蛋白质信息目录汇总数据库，提供了蛋白质准确的序列及丰富的功能注释等信息。GoPubMed在形成诱导本体时，能够提取出PubMed摘要内容中的UniProt蛋白质信息。

2.3 语义检索技术

GoPubMed实现语义检索的核心在于应用基因本体来检索和浏览PubMed。这一目标的实现有赖于两个问题的解决：一是怎样从PubMed摘要中提取GO术语；第二，怎样来构建GO的相关诱导本体。

2.3.1 术语提取

基于领域本体检索的最核心问题是本体术语与文本的映射，即术语提取。GoPubMed中存在一个基因本体的注释数据库，现已包含了3,300万本体术语的注释信息[3]，这一信息被用来进行本体术语与目标文本的匹配。

在进行术语提取前，首先需要制定术语提取原则：（1）本体术语构成树形结构，而且每个术语都包含了详细的注释信息；通常情况下，每个术语的最右边的词（即最常见的词）在自然语言处理中往往出现频率很高，但实际意义又不大（例如活性、生物合成等等），这类词也叫做停用词（stop word），归入停用词表。术语提取时应该首先忽略这样的泛词；（2）术语提取应该以首次匹配的术语（也称种子术语）作为起点。一般情况下，种子术语比较短，比较常见且概括性强，其下位术语往往描述其某一具体方面的功能和特征。然后，再从种子术语的基础上进行上下扩展，匹配到更广范围的相关术语；（3）如果不止一个术语能够与摘要内容相匹配，那么本体中相关的术语都会被提取出来，而且在形成诱导本体时维持基因本体术语之间的既有结构不变。然而，“忽略停用词”和“首先匹配‘种子术语’”这两个要求意味着尽管一些GO术语完全出现在摘要内容中，但是也有可能不能被提取出来。例如，GO术语“extracellular matrix structural constituent conferring tensile strength”，由于“strength”在停用词表中，因此，匹配时不予考虑；于是从“tensile”开始寻找匹配术语，但是，“tensile”没有出现在任何其他的GO术语中，因此，最终“extracellular matrix structural constituent conferring tensile strength”不能作为种子术语，即这一术语不能被

提取出来。导致这一问题的直接原因是GO术语及其注释信息与PubMed摘要内容的关键词字面匹配机制没有将语义纳入其中。

GoPubMed在此原则基础上进行GO术语与PubMed摘要的语义匹配。首先，将文本看作是由若干词汇组成的词序列，且按照从左到右的顺序分别赋予每个词汇编号n，n-1，…，1。这样，定义w_n，…，w_1为文本中的词汇，则s = (w_n，…，w_1)是一个长度为n的词序列。如果$n \geq i_m \geq \cdots \geq i_1 \geq 1$，那么s’ = ($w_{im}$，…，$w_{i1}$)就为s的一个子序列。一串连续的词序列就叫作一个片段（a segment）。再将PubMed摘要看作一个包含自由文本的长词序列（long word sequence），将GO术语看作一个短词序列（short word sequence）。摘要内容中的词汇（w_n，w_{n-1}，…w_1）按照由右向左的顺序，遵循迭代（循环执行，反复执行）的方法寻找与词汇w_i（i=1，…，n）相匹配的GO术语。匹配时，从摘要内容的最后一个词起，形成的片段记作“t = (w_n，…，w_1，c)”，其中“c”记为这一片段的终止词，一般情况下，“c”代表空值或者停用词表中的词。GoPubMed应用一种特殊的tokenizer算法[10]进行词序列的切分。如果词w_i是停用词表中的词，则跳过不予考虑；否则，就要通过优先匹配以词w_i开头，后接其他词汇或符号（包括空格、破折号、斜线等）的GO术语作为种子术语，当然遍历基因本体的所有层会匹配到多个种子术语，然后再向种子术语的下位或上位术语进行扩展匹配，直到遍历完整个基因本体树，匹配到所有相关的GO术语。

2.3.2 诱导本体的形成

一旦所有相关的GO术语均被提取出来，为了避免基因本体中与既定的摘要内容不相关的部分本体术语的干扰，必须建立涵盖了从摘要中提取出的所有的GO术语形成的一个最小子本体（the minimal sub-ontology）。其构建方法如下：

首先，给定本体O=(T; root; Is-a)，其中T是一个有限术语集，root ∈ T，Is-a⊆T × T，令术语t_i ∈T，1 ≤ i≤ n，且t_i Is-a t_{i+1}（1 ≤ i＜n），即t_i继承于 t_{i+1}。则p = [t_1，…，t_n]叫做从t_1到 t_n长度为n的一个路径。从自由文本中提取出相关的一套术语集T"⊆T，t"∈T"，从T"中的术语t"开始迭代查询t"的上位术语，即反复执行直到最根部，来找到包含在从T"中的术语到root的所有路径内的所有的中间术语（intermediate terms）。其次，在这一过程中，必须最大程度地考虑T中所有的术语，以使最终诱导本体结构的复杂度为O(|T|)，即保证诱导本体中术语之间的结构与基因本体术语之间的既有结构保持一致。

总之，GoPubMed通过引入“基于本体的浏览”这一概念，将领域本体作为结构化的背景知识来分层次地组织既定的语料库中的文本。例如，自由词“老年痴呆症”是与基因本体中的“脑发育”、“细胞”、“记忆”等联系在一起的。当用关键词“老年痴呆症”在GoPubMed中进行检索时，检索结果即会以GO中的“脑发育”、“细胞”、“记忆”等作为与“老年痴呆症”相关的术语通过类目分类显示出来，从而实现了检索结果的自动分类，解决了要精简检索结果这一问题。

3 SEGoPubMed语义检索机制

尽管GoPubMed已经使PubMed的检索效率得到了极大的提高，然而其检索仍然是基于布尔逻辑的关键词字面匹配，无法解决多个词汇的同义性和单个词汇的多义性的问题，无法实现完整的语义匹配；而且每一类目下的文献按照时间顺序排列，而非相关性排序。因此，研究设计基于概念（或语义）层面的匹配技术，成为又一个被关注的问题。目前，生物医学已成为语义网技术最具前途的应用领域，因为该学科领域有大量的结构化或半结构化的数据库以及知识本体[11]。此时，SEGoPubMed便应运而生。它是由美国孟菲斯大学的CVPIA（Central Valley Project Improvement Act）实验室研发的一个语义支持技术（Semantics Enabled Technique），将PubMed连接到Gene Ontology，故而得名SEGoPubMed[12]。它应用潜在语义分析（Latent Semantic Analysis，LSA）技术开创了语义敏感地浏览、连接大型文本语料库与本体的新模式。

3.1 基本思想

SEGoPubMed与GoPubMed一样，均将PubMed作为其来源数据库。然而，两者的最大区别的在于接收PubMed检索结果后，GO术语与目标文本的匹配模式以及诱导本体各类目下检索结果的排序机制。

SEGoPubMed应用LSA从一个大规模的词汇-

文本矩阵开始，通过奇异值分解（singular value decomposition，SVD）计算求出k-秩近似矩阵（词汇文本相关关系中最主要的部分），用以消减原始词汇-文本矩阵中包含的词汇之间的“斜交”噪声，更加突显出词汇与文本之间隐含的语义结构。GO术语作为查询向量，在与目标文本匹配之前将查询向量和文本向量映射到语义空间。通过语义相似度计算实现GO术语与PubMed摘要的匹配。最后语义相似度的大小用于检出结果的相关性排序。

3.2 语义检索原理

SEGoPubMed语义检索原理如图1所示。

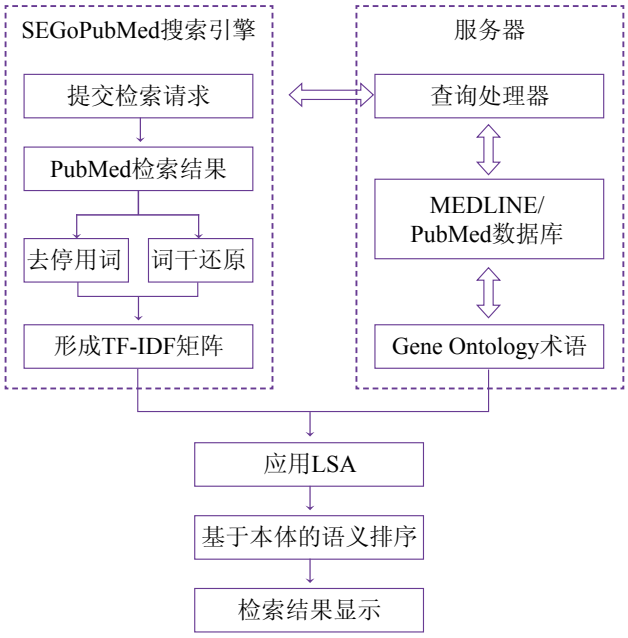

图1 SEGoPubMed语义检索原理图

3.2.1 文本预处理（Text Preprocessing）

SEGoPubMed接收到从PubMed返回的检索结果后，建立临时语料库，首先进行文本预处理，以从摘要中提取出在语义上有意义的词汇。其操作程序是：1）去除停用词；2）对英文进行词干还原；3）建立所有已提取了词干的词的“词汇库”。

这样从临时语料库中建立一个词汇-文本矩阵，将“词汇库”中所有独立的词汇作为一行，所有的文本（PubMed摘要）作为一列，词汇的权数作为矩阵的元素。权数要考虑词汇k_i在文档d_j中的频率（即词频，TF）；以及词汇k_i在文档集合中出现的文档篇数（即词汇的文档频率，DF），DF反映了词汇的区分度，DF越高表示词汇越普遍，其区分度越低，因此权数也越低。词汇的逆文档频率（Inverse DF，IDF）是DF的倒数，一般采用如下公式进行计算（N是所有文档的数目）：

IDF=log（N/DF）

通常情况下采用TF*IDF的方式计算权数，形成TF-IDF矩阵。接着是在TF-IDF矩阵的基础上应用LSA。

3.2.2 潜在语义分析（LSA）

（1）语义空间的形成及其优化

首先将TF-IDF矩阵A归一化（列归一化，即每个词汇的权数除以该词汇在所有文本中权数的最大值）。LSA通过奇异值分解将高维的TF-IDF矩阵转化为低维矩阵，分解得T，S，D^T三个矩阵，即

$A=TSD^T$

A被分解成三个矩阵的乘积，其中，T为t× r阶词汇—词汇关联矩阵，D^T为r× d阶文本—文本关联矩阵，S=diag（σ_1，σ_2，…，σ_r）为A的奇异值对角矩阵，对角线元素为奇异值，且奇异值按照递减顺序排列。

选择适当的正整数k（0<k<r且k<<min（t,N），t为词汇总数，N为文本总数），在S中仅考虑其中值最大的k个元素（即词汇、文本相关关系中最主要的部分），取S中相应的k阶对角矩阵，T中相应的k列，D中相应的k行，最终得到新的T_s（t×k阵），S_s（k×k对角矩阵），D_s（k×d阵），即

$A\approx A\hat{} = T_sS_sD_s^T$

A^即为经过优化的语义结构。T_s为词汇—词汇矩阵，D_s为文本—文本矩阵。

将每一个GO术语作为提问向量q_0，由于在上一步骤文本—词汇矩阵A已经进行了截断奇异值分解，所以对于q_0也要进行类似的处理，使其可以和文本矩阵D进行比较，用下式对q_0进行处理，即

$q=q_0^TT_SS_S^{-1}$

这样，词汇、文本、提问式三者的坐标向量构成了我们所需的潜在语义空间。

（2）文本与提问式的相关度

对提问式坐标向量q与文本矩阵D的每一行向量进行比较，可分别计算出每篇文本与GO术语间的相

关程度，即语义相似度（Semantic Similarity，简称SEsim）。最为简单常用的方法是求二者的夹角余弦值，即

$$SEsim(d_j,q)=\cos\theta_j=\frac{d_j^T q}{\sqrt{\sum_{i=1}^{k} q^2}\ \sqrt{\sum_{i=1}^{k} d_j^2}}$$

夹角的余弦值越大，则GO术语与文本的相关度越大。这一检索模型保证了每一个GO术语下能够按照其相似度值的降序排列方式输出命中的结果文档。

（3）自动赋予相似度阈值

SEGoPubMed应用自动赋予相似度阈值[3]（Automated thresholding）的技术提取与某一GO术语语义相关的PubMed摘要。凡是语义相似度值大于该阈值的PubMed摘要，都将作为检索结果提交给用户。这样，用户就可以应用本体术语按照相关性排序的方式来浏览PubMed检索结果。SEGoPubMed的“自动获取相似度阈值”的技术保证了只输出与GO术语在语义上相关的PubMed摘要，而仅有关键词字面匹配但在语义上不相关的PubMed摘要将会自动被“屏蔽”掉。

3.3 SEGoPubMed研究进展

美国孟菲斯大学应用一些引用良好（well-referenced query words）的关键词（例如“Levimisole Inhibitor”或“rab 5”）对SEGoPubMed和GoPubMed的语义检索性能进行比较。结果发现SEGoPubMed应用领域本体对PubMed进行检索时引入了语义支持；检索结果的语义相关度排序大大提高了检索效率[3]；应用的阈技术（thresholding technique）保证只提供在语义上相关的文献，证明了语义敏感技术比字符串匹配技术在将PubMed摘要与GO术语进行关联方面要优越。SEGOPubMed也能够检索到这样的文章：其中的关键词并不是孤立出现的（例如与其他术语结合在一起），而简单的关键词匹配技术是无法检索到这些文献的。然而，目前有关SEGoPubMed的研究仅局限在应用一些引用良好的关键词进行SEGoPubMed的语义检索试验阶段，而综合的、广泛的基于语义相似度的SEGoPubMed尚在研究之中。而且，现有的技术在将PubMed摘要与GO术语进行映射时也尚未实现多义词的处理，这一点可以通过应用概率隐性语义分析（Probabilistic Latent SemanticAnalysis，PLSA）技术来克服[3]。

4 结束语

语义检索作为信息检索领域的一个新的研究方向，研究人员在生物医学领域开展了颇有成效的探索。其中德国Transinsight公司的旗舰产品GoPubMed已经帮助德国600,000从事生命科学与医学研究的科学家提高领域知识搜索的效率达10-15%，由此帮助他们将更多的时间用于科学研究[13]。而正处于研发阶段的SEGoPubMed作为Gene Ontology在MENLINE这样一个超大型数据库上的语义层次的尝试，其产生本身确实是一个创举。PubMed、GoPubMed、SEGoPubMed代表了信息检索从基于叙词表发展到基于本体和语义网的知识工程领域的巨大进步。这对于生物医学领域研究人员洞悉生命科学与医学发展脉络，开展创造性的基础研究具有重要意义。其开发应用的成功也为其他领域的智能语义检索研究提供了值得借鉴的样板。

参考文献

[1] PubMed Home [EB/OL]. [2009-03-26]. http://www.ncbi.nlm.nih.gov/sites/entrez/.
[2] 代涛,高军. 卫生政策研究信息资源与检索[M]. 北京:人民卫生出版社,2008:50-51.
[3] VANTERU B C, SHAIK J S, YEASIN M. Semantically linking and browsing PubMed abstracts with gene ontology[J]. BMC Genomics, 2008, 9(Suppl 1): S10.
[4] 杨俊柯,杨贯中,杨建学. 基于语义模型的信息检索机制研究[J]. 计算机工程,2006,32(12):212-214.
[5] Transinsight GmbH - think the impossible [EB/OL].[2009-03-29]. http://www.transinsight.com/.
[6] ASHBURNER M, BALL C, BLAKE J, et al. Gene Ontology: Tool for the unification of biology [J]. Nature Genetics, 2000, 25 (1): 25-29.
[7] Gene Ontology Home [EB/OL]. [2009-04-06]. http://www.geneontology.org/.
[8] United States. National Library of Medicine. Medical Subject Headings (MeSH®) [EB/OL]. [2009-01-17]. http://www.nlm.nih.gov/pubs/factsheets/mesh.html.
[9] 陈龙,朱化彬,沙里金. 蛋白质组学数据库建设的研究进展[J]. 畜牧与兽医,2008,40(8):102-104.
[10] DOMS A, SCHROEDER M. GoPubMed:exploring PubMed with the gene ontology [J] .Nucleic Acids Res, 2005,33, 33 (Web Server issue): W783-6.
[11] DOMS A, JAKONIENE V, LAMBRIX P, etc. Ontologies and Text Mining as a Basis for a Semantic Web for the Life Sciences [M]. Springer Berlin / Heidelberg, 2006.
[12] ROUCHKA E C, KRUSHKAL J, GOLDOWITZ D. Proceedings of the Seventh Annual UT-ORNL-KBRIN Bioinformatics Summit 2008 [C]. BMC Bioinformatics,

9(Suppl 7):11.
[13] ALVERS M R. GoPubMed – Knowledge based search [J].Technology and Health Care, 2007 (15): 295-296.

作者简介

杜建（1986-），男，硕士研究生，研究方向为医学信息管理。通讯地址：湖北省武汉市航空路13号华中科技大学同济医学院医药信息系 430030
张士靖（1958-），女，副教授，硕士生导师。主要研究方向为情报语言学、医学信息检索、医学信息素质教育。通讯地址：同上。E-mail: zhangsj9999@163.com

Research of Biomedical Semantic Retrieval Based on Domain Ontology——A Case Study of GoPubMed and SEGoPubMed

Du Jian, Zhang Shijing / The Department of Medical Information of TongJi Medical College of HuaZhong University of Science and Technology, Wuhan, 430030

Abstract: Domain ontology based semantic retrieval of the large-scale text corpus has become a hot area of information retrieval research in recent years. GoPubMed and SEGoPubMed, as an outstanding representative of this research area, provide good solutions to the existing problems of PubMed. Our deep studies on semantic search mechanism of GoPubMed and SEGoPubMed have shown that GoPubMed applies Semantic Web-based classification instruments—GO and MeSH, measures and corresponds the PubMed search results to the classification of the corresponding navigation directory so that users can use the navigation to quickly find the literature they need, however，it is still based on the string matching mechanism of Boolean logic model. SEGoPubMed,which performaces better than GoPubMed, applies Latent Semantic Analysis (LSA) technology to create a new paradigm to semantic-sensitive ontology based browsing and linking of large corpus to ontologies. The analysis suggests that the SEGoPubmed performs better than the classic GoPubmed as it incorporates semantics. SEGoPubMed has a brighter future.

Keywords: Domain Ontology, Gene Ontology, MeSH, GoPubMed, SEGoPubMed, Latent semantic analysis

（收稿日期：2009-06-26）

业界动态

全球最早大屏电子书厂商iRex申请破产

iRex算是电子书阅读器行业的先锋之一，也是全球最早推出大屏幕电子阅读器的厂商，早在2006年就推出了可以手写的iRex iLiad 电子书，而国外媒体报道，iRex 因为「财务困难」宣告破产。

据说最主要的问题，是来自于 DR800 电子书在FCC那里迟迟过不了关，导致iRex错过了去年年底的圣诞节购物潮，但或许它贵得要死的价格也不符合一般人对电子书阅读器的期望吧。接下来就看 iRex有没有机会重整啦！

iRex CEO汉斯 · 布隆斯（Hans Brons）表示，该公司今年2月推出的电子阅读器iRexDR800SG销售情况未达预期，该公司已经用光了“运营资本”。

DR800SG的售价为399美元，在推出之初曾由于其开放模式而备受关注。这款产品使发行商更好地掌控内容的发布。与亚马逊Kindle不同，发行商可以对电子书自由定价。为推广这款产品，iRex此前曾与百思买和Barnes&Noble签订合作销售协议。

在国内，一直作为国内总代而存在的“广州东豪”也神秘消失，东豪原来的网站irexchina.com也打不开了。但是在两个月以前，国内还在不少经销商在销售“增强版”的iliad一代，“广州东豪”的神秘消失应该是早已收到iRex破产的消息，国内众多iLiad用户将面临着售后服务问题。

在国外有破产程序，用户权益一般会得到保障，而国内厂商则通常是关门跑人，这应该引起行业和用户的警惕，大批的厂商挤入电子阅读器行业，真正拥有核心技术，能长久做下去的也许并不多，未来会有更多的无竞争力的厂商倒闭，用户应尽量选择有实力的厂商产品。

来源：http://news.xinhuanet.com/shuhua/2010-06/14/content_13669350.htm 查询日期：2010-06-15