DOI：10.13255/j.cnki.jusst.2011.01.007
上海理工大学学报
第33卷 第1期 J. University of Shanghai for Science and Technology Vol.33 No.1 2011

文章编号：1007－6735(2011)01－0018－06

思维与智慧科学及工程

尹红风[1]， 戴汝为[2]
(1.西南交通大学 信息科学与技术学院，成都 610030；
2.中国科学院 自动化研究所 复杂系统与智能科学重点实验室，北京 100190)

摘要：讨论了钱学森的思维科学、开放复杂巨系统和大成智慧的理论，互联网和云计算的发展，可用这些研究和理论建立类似人的世界知识库，用于语义搜索引擎.

关键词：思维科学；开放复杂巨系统；大成智慧；语义搜索引擎
中图分类号：N 94 **文献标志码**：A

Noetic and intelligence sciences and engineering

YIN Hong-feng[1]， DAI Ru-wei[2]
(1. *College of Information Science and Technology，South East Jaotong University，Chengdu* 610030，*China*；
2. *Key Laboratory of Complex Systems and Intelligence Science Institute of Automation，Chinese Academy of Science，Beijing* 100190，*China*)

Abstract：Qian Xuesen's noetic science，open giant complex system theories and metasynthetic engineering were discussed. With the development of Internet and cloud computing，the research and theories can be applied to build a human-like world's knowledge system for semantic search engine.

Key words：*noetic science*；*open giant complex system*；*metasynthetic engineering*；*semantic search engine*.

从20世纪80年代起，科学大师钱学森提出思维科学、开放复杂巨系统、人-机共建的智能系统和综合集成的大成智慧等一系列的思想和理论，我们与他一起开展了这些研究，钱学森当时预言：这是科学的革命，必将带来技术的革命. 今天可以更清楚地认识到钱学森开创的思维与智慧科学革命，这是中国第一次在重大科学问题上领先突破. 本世纪伊始，认识到思维与智慧科学思想和理论正是新一代语义智能搜索引擎的理论基础，新一代搜索引擎就是智能计算机，信息技术的新发展使得今天完全可以在工程上实现这些理论，从而开启新的知识技术革命.

1 思维科学的研究与发展

1.1 人工智能的困境

物质的本质，宇宙的起源，生命的本质和智能的产生是人类科学所面临的四大挑战. 国际上对智能的研究主要是用人工智能的方法. 1956年，第一次人工智能研讨会在美国的达特茅斯(Dartmouth)大学举行，J. McCarthy，H. Simon等倡议开展人类思维活动

收稿日期：2010－11－25
作者简介：尹红风(1964－)男，教授. 研究方向：数据挖掘、人工智能、机器学习和语义智能搜索.
E-mail：hongfeng@yebol.com

规律的研究，并给予“人工智能”的命名，标志着人工智能学科的诞生．人工智能的实现主要是基于逻辑符号处理，并且主要以机器模拟人的智能为主，但其方法论和目标存在着问题，为后来的研究者埋下了束缚思想的桎梏．对游戏、下棋和机器定理证明等问题容易解决，1958 年 H．Simon 曾乐观的预计：10 年之内计算机将成为世界象棋冠军、发现并证明重要数学定理、谱写出优秀的乐曲，到 2000 年，机器的智能将超过人……．但是在自然语言理解和机器翻译研究则遇到瓶颈．80 年代日本提出第五代智能计算机计划，主要是提高逻辑运算的能力．第五代机计划的失败是对传统的人工智能研究的另一大的冲击．

1.2　思维科学

对智能本质的研究，科学大师钱学森的思维科学开创了新的科学革命，钱学森在 20 世纪 50 年代就开始思考思维科学的研究，20 世纪 80 年代，钱学森提出人的思维是有规律的，可以用科学的方法研究，思维科学是可以成立的，并撰写了著名的《关于思维科学》一文[1]，文中指出：从广泛的意义上讲，思维当然有规律，因为思维也是一种客观现象，而一切客观的东西及其运动都有自己的规律，思维当然也不例外．可以先从思维是人的中枢神经系统，特别是大脑受外界各种刺激而引起的这一点看．外界各种刺激又是客观世界变化和运动的产物，这些变化和运动是遵循客观世界规律的，即自然界的和社会的规律，所以外界各种刺激也是有自己的规律，而不是无缘无故无章可循的．这样，人的中枢神经系统大脑的活动也就当然要有规律，人的思维要有规律．思维科学只研究思维的规律和方法．

钱学森进一步指出“思维”可以分成抽象(逻辑)思维、形象(直感)思维和灵感(顿悟)思维 3 个部分．特别强调要在“形象思维”研究方面有所突破．钱学森先生还认为计算机模拟对研究人的思维有重要的启发，计算机模拟技术是研究思维的有效工具．

钱学森先生的思维科学也得到了人工智能之父、诺贝尔经济学奖和计算机图灵奖获得者司马贺(Herbet Simon)的高度关注，他写信给钱学森，希望能和钱学森直接探讨思维科学的问题，并认为可以和钱学森共同树立一面旗帜．可惜由于各种原因，两位东西方科学大师没有能够直接对话．

钱学森认为，思维科学的研究将孕育新的科学革命，另一方面，思维科学的研究又会推动智能机的发展，肯定又将是一场技术革命．

当时用思维科学的理论来分析日本的第五代计算机计划，就认识到这是一个失败的计划，因为它的架构中没有模拟形象思维的功能．

思维科学开辟了新的正确的智能研究方向，是发展智能机的理论基础．把钱学森思维科学的思想深入发展成科学的理论和实现，写了《论思维与模拟智能》一文[2]，建立了一个思维的结构模型，详细描述了形象思维、逻辑思维和其对应的存储、运算之间的关系，更进一步实现了形象思维的联想记忆数学模型和人工神经元网络的模拟[3]．钱学森和我们进行深入探讨并对我们的工作给出很高的期望[4]．

2　从思维科学到智慧科学

2.1　开放的复杂巨系统

对于思维科学的进一步探讨，钱学森在 1989 年 8 月 24 日给的信中指出[5]：“作为物质系统如何形容人脑？认为应该用系统学的概念，人脑是由几万亿脑细胞组成的开放复杂巨系统”．钱学森在 20 世纪 90 年代初进一步发展为开放复杂巨系统理论[6]，认为开放的复杂巨系统的主要性质可以概括为：

a．开放性——系统对象及其子系统与环境之间有物质、能量、信息的交换；

b．复杂性——系统中子系统的种类繁多，子系统之间存在多种形式、多种层次的交互作用；

c．进化与涌现性——系统中子系统或基本单元之间的交互作用，从整体上演化、进化出一些独特的、新的性质，如通过自组织方式形成某种模式；

d．层次性——系统部件与功能上具有层次关系；

e．巨量性——数目极其巨大．

互联网正是一个“开放的复杂智能巨系统”：

a．巨量性——已经拥有数千亿的网页，数十亿的网民，数亿的关键词概念；

b．复杂性——互联网包括各种不同的系统，不同的行业，不同的功用；

c．开放性——用户系统、网页系统之间总是在互相作用，交换信息；

d．进化与涌现性——这些元素又互相关联，这些元素之间关系也是不断变化的，人的参与更把这些元素组织成有意义的模式；

e．层次性——概念之间不仅相关，而且有各种层次，网页也包含许多层次．

2.2　人机共建的智能系统

钱学森在 1989 年 8 月 24 日的信中还指出[6]：

“搞模拟智能的起步该在什么地方,如何从人机结合一步一步的提高?”1991年4月18日更明确指出:“智能系统是非常重要的,是国家大事,关系到下一世纪我们国家的地位.如果在这个问题上有所突破,将有深远的影响.要研究的问题不是智能机,而是人与机器相结合的智能系统.不能把人排除在外,应是一个人-机智能系统.”

2.3 综合集成的大成智慧

钱学森的大成智慧思想是把人的思维、思维的成果、人的知识、经验和智慧以及各种情报、资料、信息集成起来[7].顾名思义,称为“大成智慧工程(Metasynthetic Engineering)”.构思是把今天世界上千百万人的聪明才智和智慧都综合起来.

这样则把智能的研究的方向从人工地模拟智能的功能转变为研究人的智能原理,从个体转变为社会的智慧,从简单算法到复杂巨系统,从以机器为主到以人为主、人-机结合的的智能系统.

3 技术新浪潮

3.1 信息技术革命和发展极限

计算机的发明给人类带来了信息技术和信息革命,互联网的发展将信息革命推向新的高潮,信息存储、运算和通讯能力都成指数性增长,人们同时也面临许多垃圾、有害、虚假等信息,现有的信息技术已使人无法有效使用已有的信息,信息技术革命已到了尾声.

以信息检索理论为原理的搜索引擎是目前主要的信息寻找方法,它主要是通过网络蜘蛛尽可能搜集互联网网页,然后用超链分析等方法给出网页排名,再用关键词来索引所有的网页,最后对用户输入的关键词,搜索引擎从索引数据库中找到匹配该关键词的网页提供给用户.搜索引擎通常能够涵盖非常大的互联网范围,但是经常返回大量的低质量网页.尽管过去几年里在搜索引擎技术和系统上有许多改进,但是人们搜索网上信息时还经常有很大的挫折感,很多时候,想要的信息不能够找到或者需要花很多时间才能找到,给出的网页的数量通常也很大,并且只能够给那些它包含搜索词的网页.另外,现在的搜索引擎对所有的人几乎给出同样的搜索结果.虽然过去十年互联网发生巨大的变化,但搜索引擎还是和十年前几乎相同.

3.2 新技术浪潮

近几年,终端设备如智能手机、平板电脑和电子书等迅猛发展,特点是小屏幕、移动、联网和个性化.通讯、计算机和媒体的结合越来越密切.

在后台,云计算是计算平台的革命,通过Hadoop开放平台实现的Map/Reduce算法,可以用数万台机器来完成一项工作,几乎有无限的计算、存储和通讯能力.并且Amazon AWS等提供了硬件服务.可以以低价格、迅速、灵活地租用.在内容方面,用户产生的内容急剧增加:如博客、微搏、社交网络等.视频、图象等多媒体内容也越来越重要.

互联网的用户大规模增加,中国已有4亿多互联网用户和将近4亿移动互联网用户.

而这些技术浪潮还主要是硬件和环境的改变,需要通过一个新的系统才能把这些资源有机地整合起来,最大发挥新技术的潜力,从而转化为新的技术革命.钱学森的思想和理论正是这场新技术革命的核心和基础,而这些新的计算、设备、通讯、互联网和媒体的新发展也为实现钱学森的大成智慧工程提供了必要的条件.

4 大成智慧工程的实现

4.1 信息与知识

21世纪伊始,我们认识到钱学森的思维、智慧科学思想和理论正是新一代语义智能搜索引擎的理论基础,新一代搜索引擎就是智能计算机[8].其目标是要建立类似人的世界知识库,从而可以提供基于知识的搜索,或者说是知识引擎.只有像人一样,理解所有的信息,将巨大的信息转变成有用的知识,才能最好的利用信息,这将开启从信息技术向知识技术的巨大转变,从以数据为中心向以人为中心的转变.就探索这些技术的实现,克服算法和工程方面许多难题.

那么信息和知识之间的主要区别是什么呢?表1列出了信息与知识的比较.

表1 信息与知识的比较

Tab.1 The comparisons of information and knowledge

项目	信息	知识
范围	孤立的、局部的、线性的	关联的、综合的、整体的、网络的
系统	静态的、死的	动态的、活的、变化的、开放的
中心	数据	人
本质	符号化	个性化、概念化、有意义
结构	无结构、无序、冗余	分类的、有层次的、有结构的、有序的
增长	爆炸性	线性
处理	符号、逻辑	思维、联想
两者关系	是知识的外在表达	是信息的内在转换

人工生成的知识系统，如维基百科、网页目录等等很好建立，但这些系统尽管参与者众多但容纳的词条到底有限，只有区区几百万条．目前国际上语义搜索引擎还是处在概念化阶段，其它的语义搜索引擎如 Wolfram Alpha、Hakia、Powerset、Maholo 等只能在较少的领域或较小范围内搜索．最近，Freebase 和 DBpedia 已经把大量的网上信息结构化，从而建立关键词之间的可以用语言描述的关系，我们则用算法计算出关键词之间的联系的数字强度关系，从而可以对知识库的结果进行排序，两者结合起来则可提供用户搜索更精确的、更直接的答案．目前，DBpedia 知识库已经建立了关于 290 万事物的 4 亿 8 千万条信息．这样关于这些上百万的事物的上亿的询问，就可以给出直接答案．对中文的这样详细描述关键词之间关系的知识库，还需要建立．

4.2　建立世界知识库

我们的目标是建立任何事物的知识库，从而可以对大部分的搜索，都可以给出直接答案．那么，怎样建立一个这样的人工的开放复杂巨系统？需要应用钱学森的人机共建的综合集成理论．目前，互联网提供了实现大成智慧的几乎所有必要的条件，网上有几乎人类所有的知识、数据、资料、信息和巨大的用户，但是都是分散的、无组织的，我们则可以把这些同各种算法、系统、技术和设备集成起来，运用云计算的巨大能力，构建一个海量的知识库和智慧平台，从而可以提供各样的智慧服务．图 1 显示建立海量的知识库的综合集成方法．

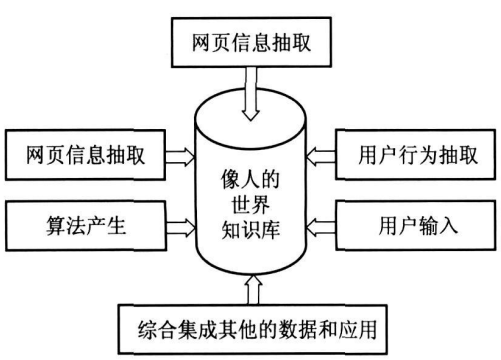

图 1　建立海量世界知识库的综合集成方法

Fig．1　Metasynthesis method for building massive world's knowlege system

4.2.1　集成信息、数据和系统

首先可以从大量的互联网页中抽取有用的、结构化的信息，对所有的网页都可抽取重要的链接、关键词信息，对某些特殊的领域和主要的网站，则可抽取更加精确和结构化的信息，如地址、电话、电影、图书、生日等．这样就可把网上的信息转化为知识，这些知识使得智能搜索可以回答一些经过推理、综合才能回答的问题，如一个人的年龄，某个市的主要医生等问题．这些问题是传统的搜索引擎所不能解决的．

互联网上还有许多公司的专业知识库和数据，如天气、股票、旅游等，我们则可把这些数据和系统直接集成到建立的智慧平台里．

4.2.2　集成人的智慧

人脑也是一个复杂巨系统，有超过 100 亿神经细胞，云计算技术的发展可建立一个人工的这样大规模的复杂巨系统，如果每台服务器可以处理 100 万单元的信息，那么 1 万台务器组成的云计算则可以处理 100 亿单元的信息，相当于人脑的运算能力，因此云计算可使有和人脑同等量级的运算能力．因此，今天能够在技术上实现开放复杂巨系统，这为进一步定量研究开放复杂巨系统理论提供了实验基础．同时通过模拟也是了解、认识复杂巨系统一个重要途径．

人使用互联网行为如搜索的词、点击的网页、浏览的网页等包含了大量信息，可以用算法处理和分析，从而得到集体的智慧，其结果可以用于内容、关键词推荐等．对个人的行为分析、处理则可为每个用户建立知识库，提供个性化的服务和搜索．例如统计所有人搜索词的频率和个人搜索词的频率，则可用于建立高效的、个性化的输入法．

数亿用户的直接输入是知识库的重要来源，象百科、复杂问题解答、博客等已经是互联网重要内容来源，用户的知识是用人的智能解决精确的问题和复杂问题．目前这些信息还不是结构化的信息，我们则可以设计结构化的界面，从而得到结构化的信息，则可以使用户输入的信息的搜索和使用的功效大大增加．

数亿互联网用户也可以看作巨大计算和智力资源，虽然每个人运算速度不快、记忆有限，但是几亿的用户的计算量积聚起来可以是巨量的，像图像识别、语音识别、机器翻译、复杂问题回答等，机器是无法和人相比的．因此，这是一个以人为主，人-机结合的系统．

机器是要把所有人的智慧综合集成起来、把其潜力发掘出来．

4.2.3　集成自然语言处理

机器算法可以处理上万亿条词目，自动产生知

识.到目前为止,人工生成的知识库与机器生成知识库之间主要的区别在于后者不如前者精确.自然语言处理最终可以用机器把大部分网页里的文字信息转化为知识.这还需要相当长的时间研究才能实现,但是我们可以一步一步的来实现这个目标,先理解一些简单的问题,抽取一部分知识丰富知识库,或对一些特定的领域处理,逐步扩大到多较复杂的问题和多领域.另外通过海量知识库提高对网页自然语言理解的能力,从而抽取更多的知识丰富知识库.

4.2.4 集成数据挖掘结果

互联网上早就产生海量数据,但是几年前,分析和处理海量数据是一个巨大的工程,往往要耗费数十人,数个月甚至一、两年时间.研究数据挖掘算法大部分时间是用在产生数据上.云计算提供了方便、快速处理海量数据的平台.可把产生数据的时间从几个月缩小到几天、甚至几个小时,这是继个人计算机后计算平台的一次革命.

海量数据还使得许多过去算法如机器翻译、图象分类、自然语言处理等都会有新的方法和结果的突破,把过去一些规则、学习和分析的方法变为海量样本的搜索和比对.

怎样从海量数据中用数据挖掘算法产生知识、自动产生分类、聚类等结果?互联网数据有以下特点:

特点1 数亿至数万亿条以上信息,如个人行为信息、网页信息、关键词信息等.

特点2 数据特征维数可达百万以上,如对文本,如果每个关键词都可看作一个特征.数据非常稀疏.

特点3 可以来自多个数据源,如人行为数据有:搜索词、浏览的网页、看到和点击的广告、购买的产品等.

因为数据挖掘一般都是非常大的工程项目,并且有很重要的商业目标,涉及许多人和各种资源,即使是在工业界,成功的也是很少.数据挖掘项目的成功取决于如下重要因素:

因素1 选择数据.因为现代信息技术可以产生巨量的数据,有不同的数据源,但是要用什么样的数据参与挖掘?数据与目标的相关性如何?成本如何?有时数据量巨大但含的有效信息较少,有时数据极为有效但量太少.怎样取舍?需要事先有定性的分析和判断,这往往需要很多数据挖掘的经验和专业知识的经验.同时也需要先用少量数据进行分析和验证大的设想.

因素2 探索数据.当选择好要用的数据后,还需对数据本身进行认真仔细观察、分析、探索、统计结果和每一特征的分布等,研究数据的可靠性和稳定性等,及早发现数据可能存在的问题.并且数据还需要进行变换以符合算法的要求.从数据中发现新的思想.

因素3 产生训练样本.需要从海量数据中选择一定量的学习数据和评价数据的进行建模,选择多少和选择哪些样本数据对模型的结果有很大影响.

因素4 运用算法.通常各种数据挖掘的算法得出的结果差别并不是特别大,对许多实际问题,结果如能满足客户的主要要求,我们主张尽可能用简单的算法,如线性回归算法(Linear Regression)或Logistic Regression,KNN, 神经元网络算法等.

因素5 熟悉运算和系统平台.要了解云运算Hadoop平台和其他的相关的系统,才能有效地产生数据,把训练好的模型集成到实际运行的系统中,要考虑和实现运算速度、系统集成等要求.

因素6 了解市场需求.另外,还需了解市场的实际效果和需求,不断改进,设计和开发新一代产品.

4.3 知识库的管理

通过各种方法产生海量知识后,还需要对这些知识有效地管理,主要有以下几个方面:

a. 知识的更新.对从网页中抽取的知识要跟据网页内容变化的频率自动下载更新.对数据挖掘算法和自然语言处理算法产生的知识要根据需要每个星期或每天运行算法.也可以设置界面让用户直接更新.

b. 知识的排序.为了能够对海量的知识有效地查寻,需要进行排序,对每条知识根据其来源、用户关注度、搜索频率以及内容的大小等打分,将来也可通过学习算法打分.根据分数可对搜索结果进行排序.

c. 知识的歧义和同义.对一个名称可能有不同的含义,如苹果可以是公司或水果,同一人名可以是不同的人.另外,对于同一内容也可以有不同的名称,如北大和北京大学多是指同一内容.

d. 知识推理.综合多条知识或数据根据一定的规则、科学公式或训练的数学模型给出结果,如从生日给出年龄,数学运算.

4.4 海量知识库和智慧平台的应用

当建立了这样的海量知识库和智慧平台后,就

可以用于不同的方面如图 2 所示，首先可以提供快速而准确的语义智能搜索服务．并且可以自动产生数亿的高质量的内容．也可以进行自动内容分析，并最终实现人机自然对话．

由于在云计算的平台上实现优质的服务，有足够的存储空间、计算能力和网路带宽满足系统的需要，因为云计算是根据实际的用量来收费，这也大大降低了费用．

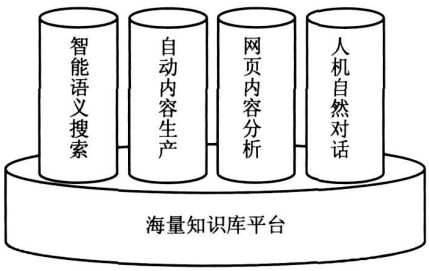

图 2　海量知识库和智慧平台的各种应用

Fig．2　Applications of massive world's knowledge and intellgence platform

5　结束语

可以看到，钱学森晚年的思维科学、开放复杂巨系统、人机共建的智能系统和综和集成的大成智慧等研究对人工智能、计算机科学、信息科学等的新发展有着奠基性的指导意义，是中国第一次在重大科学问题上领先突破．目前互联网终端和云计算技术的发展终于可以实现他的这些的理论和思想，建立海量的知识库和智慧平台．这将是一个用云计算集成几十亿终端、和几乎所有人类信息和数据以及几十亿的网民行为和智慧的开放复杂的海量系统，从而带来从信息到知识的技术革命．钱学森的研究将对人类的思想、科学和技术作出伟大的贡献．

参考文献：

[1]　钱学森．关于思维科学[M]．上海：上海人民出版社，1986．

[2]　尹红风，戴汝为．论思维及模拟智能[J]．计算机研究与发展，1990(4)：1－16．

[3]　尹红风，戴汝为．一种联想记忆模型及附加节点方法[J]．计算机学报，1990，13(5)：331－340．

[4]　钱学森．致戴汝为——1989 年 5 月 14 日[M]//涂元季．钱学森书信(4)．北京：国防工业出版社，2010：484－487．

[5]　钱学森．致戴汝为——1989 年 8 月 24 日[M]//涂元季．钱学森书信(5)．北京：国防工业出版社，2010：23－26．

[6]　钱学森，于景元，戴汝为．一个科学新领域——开放的复杂巨系统及其方法论[J]．自然杂志，1990(1)：1－10．

[7]　戴汝为．钱学森论大成智慧工程[J]．中国工程科学，2001，3(2)：14－20．

[8]　戴汝为，尹红风．从思维科学到知识技术革命[N]．科学时报，2009－12－29(A2)．

（责任编辑：巩红晓）

·下期发表论文摘要预告·

神经系统中的复杂性研究

方福康

（北京师范大学，北京 100875）

神经系统面临的一个基本科学问题是要回答神经系统中微观到宏观的机制．微观运动形式以神经元突触活动为基础，宏观行为包括视觉、听觉等基本认知功能，以及大脑思维的高级认知．微观与宏观的差异是巨大的，其过渡的形式困扰着神经科学，迄今未获解决．系统科学复杂性研究的出现，为这个科学问题提供了新的角度和方法．该文讨论了神经系统信息传递和转换中的突变行为，通过视知觉、概念形成、记忆等研究案例，阐明了复杂性研究对神经系统的处理方案与结果．文章也讨论了学习过程中几种最基本的突变形式．