语言计算:信息科学技术中长期发展的战略制高点

清华大学智能技术与系统国家重点实验室
孙茂松

一、 基于语义的内容计算

随着互联网以及大规模数据存储体系的迅猛发展，人类已经进入名副其实的海量信息时代。例如，著名的搜索引擎 Google 的检索范围已达 80 多亿张网页，允许对近三十种语言进行搜索（包括英语、主要欧洲国家语言、日语、中文简繁体、朝鲜语等）。人类知识更新的步伐日新月异。据激光打印机发明人 Gary Starkweather 博士称：在 1750～1950 年中，知识增长的速度是 150 年翻一番，而 1950～1960 年间，10 年就翻了一番，1960～1992 年间，翻番时间已缩短到 5 年。期望到 2020 年，信息量每 73 天就将翻一番。绝大多数新产生出来的信息都是数字化的，同时旧的信息也正在通过大型的数字图书馆计划不断地在被数字化中。可以设想，在不远的将来，互联网上将集聚人类有史以来创造的几乎全部知识。

然而，拥有海量数据仅仅意味着人类拥有全面、深入、方便地驾驭这些海量数据中所蕴涵知识的潜在可能性，但可能性与现实性有天壤之别。现实状况是：目前对海量数据的操作主要还在信息检索阶段，根本谈不上构建于其上的知识组织、总结及分析。即使是信息检索这个比较初级的任务，效果也很不理想：TREC 2004 Terabyte Track 的测试结果显示，文本信息检索的最高精度不超过 30%。而对声音、图象、视像等的搜索能力就更差了。就目前状况而言，互联网这个知识海洋颇像虚拟世界中巨大无比的“黑洞”，大多数宝物都被默默地埋藏于幽深的海底难见天日，而我们却缺乏有效手段实现随心所欲的“大海捞针”，只好无奈地“望洋兴叹”。人类正面临着一种前所未有的尴尬与困惑的局面：对数字信息利用的有效率极其低下。换个形象的说法, 互联网象个大茶壶,它的壶体正在急剧膨胀,颇有“醉里乾坤大, 壶中日月长”的味道,但茶壶嘴几乎没有扩张,虽然大肚能容,有货却倒不出来。

必须指出，计算机的运算速度、磁盘容量、存取效率、网络带宽等因素与解决这个问题并无实质性关系（著名的摩尔定律指出，计算机的性能每 18 个月翻一番。目前的发展实际超越了摩尔定律，如 3 年内图形处理能力提高了 100 倍，网络带宽增加了 64 倍）。彻底扭转此被动局面的唯一途径是，信息处理必须跨越到基于语义的内容计算。

这一跨越在信息处理的研究与应用两大方面都将是无与伦比的，一旦得以完成，将会导致信息技术出现一场全新的革命，推动人类从虚拟世界的必然王国走进自由王国，其重大意义无论怎么讲都不过分，经济效益和社会效益不可估量：

（1）科学意义：实现以信息为中心的计算（Information-centric Computing）。放大人类的智能，而非简单地放大人类的工具。

（2）国家基础设施建设：从 Web 走向互联网发明人暨 W3C 主席 Tim Berbers Lee1998 年提出的语义 Web(虽然笔者认为, 在中近期实现严格意义的语义 Web 近乎天方夜谈,但其变体,如面向特定应用的小型语义 Web 却是可能的)，提升信息的质量与系统性，实现知识的有效组织与利用。

（3）国家经济建设：建设与工程体系相配合的、以“软科学”为特征的非工程体系，提供全面、强大的决策支持。

（4）国家安全：敏感信息的准确检测与过滤（例如军事、政治敏感信息）。目前基于 IP 地址及基于关键词匹配的策略只能是权宜之计，防不胜防。

（5）人民生活质量与文化素质的提高：网络的各种个性化服务及按需服务。

（6）网络色情的围堵：有效制止其恶性泛滥（已成为网络上的首要公害）。

虽然要圆“基于语义的内容计算”之梦，人类还需要走非常漫长的路,但在这个圆梦之路的不同阶段所产生的一些阶段性重要成果，仍足以促使信息技术发生深刻变革及带动相关产业的升级。

由于自然语言文本占据了互联网的大半河山, 同时,在可预期的将来,对声音、影像、图片的检索仍将严重依赖自然语言分析技术(正如近两年 Google 推出的图象与视像搜索引擎所做的那样),语言计算的重要性也就不言而喻了。可以预期,它将无可避免地成为信息科学技术中长期发展的战略制高点。

二、 目前的主要任务

语言计算是一项长期的艰巨任务，不可能一蹴而就。那么,在现阶段,我们应该抓的主要任务有那些呢?笔者认为,以下诸多方面的研究应该成为我们关心的焦点。

（1）在人工智能、机器学习、数学、语言学等理论交叉指导下，进行面向超大规模文本等真实复杂环境的方法与原型研究。尤其要注意研究算法在这一条件下的性质。

（2）面向互联网的汉语自动分词研究

英文的信息处理一起步就是在词平面上。而中文信息处理起步是在字平面上。现有的中文搜索引擎，虽然几乎使都用了汉语分词系统，但由于分词系统的性能存在严重缺陷，导致检索性能不佳，更堪担忧的是，中文搜索引擎将无法向更高级的形态发展。许嘉璐先生曾一针见血地指出：“到目前为止，中文信息处理基本上还停留在‘字处理阶段’，也就是说计算机对汉语的‘认知’是一个字一个字地进行”“如果我们说得‘宽宏’一些，最多可以说现在是处在‘字和词处理之间’阶段”“中文信息处理技术虽然在有些方面有所进步，但至今还没有跨上‘语言处理’这个台阶”。要从字平面跨越到词平面，汉语自动分词是必由之途。观察表明,现有的分词系统对互联网文本的处理能力远远不够。这个貌似简单的任务其实十分困难，不以大工程的态度对待，断无成功之理：

- 建立“信息处理用现代汉语通用分词词表”,与国家标准“信息处理用现代汉语分词规范” 相互衔接。这个通用词表将成为构造语义 Web 所需的通用 ontology 的基础。
- 建立各个主要应用领域的分词词表（词数当在数百万级），并制订相关规范。这些领域分词词表将成为各领域 ontologies 的基础。
- “来自互联网”: 在通用词表与领域词表的支持下，以互联网上的中文文本集合为基本对象，进行汉语分词歧义等的大规模调查，据之设计有效的分词歧义消解算法,并进行新词汇自动发现的研究。
- “面向互联网”: 实现一个可驾御互联网的实用型汉语自动分词系统。研究当分词必然存在一定错误率的条件下中文搜索引擎设计的健壮性问题。

（3）应用驱动的浅层句法分析技术的研究。

（4）借鉴 WordNet 与 HowNet，进行大规模汉语语义资源的整合与建设。并且以之为基础，进行汉语语义计算的研究。

（5）词法、句法、语义一体化的汉语分析模型的研究。

（6）进行领域 ontology 的研究，并建立一个示范性 ontology。制订相关的标准。

（7）研究在海量文本中自动发现词与词之间关系的算法。

（8）研究高精度的汉语文本自动分类算法,建立 Web 逻辑地图。

（9）将自然语言处理、OCR、语音识别等技术融合于基于内容的图像、视像处理研究中，以显著提高图像和视像的智能化处理能力。

（10）完成对文本、声音、图像和视像均具有很强判断能力的关键性应用系统（典型如色情和军事、政治敏感信息的自动过滤）。

（11）促进大规模语言计算资源共享平台与机制的建设。

（12）将上述成果集成起来，设计并实现实用型工具软件,可以将任意一个普通网站经过若干步深层次处理后自动转换成一个智能型网站，从而被赋予一定的知识管理能力。

（13）建立并完善我们自己的搜索引擎,与 Google 抗衡。

（14）在内容计算的基础上，研发各类知识服务系统,如基于 Web 的预警系统。

三、 中文信息处理与中华文化

中国已成为世界上仅次于美国的第二大网络大国。据中国互联网络信息中心《第十五次中国互联网络发展状况分析报告》统计，截止到 2004 年 12 月 31 日，我国网民数量已经达到 9400 万，上网计算机总数已达 4160 万台,WWW 站点数约 668,900 个。中国大陆 IP 地址总数已达 59,945,728 个。中文网上资源呈雪崩式发展的态势。

于是, 语言计算无可避免地被赋加了更多一层的特殊意义：面向中国人的语言计算, 以确保国民掌控及利用海量中文信息的能力，同时，使中华文化能够借重这种强有力的技术手段，在全球网络一体化，多文明、多文化共存乃至激烈碰撞的考验下，岿然屹立于世界文化之林，并且历久弥新，发扬光大。令人深思的一个例子是,2004 年 12 月，Google 宣布将对纽约公共图书馆及四家知名大学图书馆（牛津大学、哈佛大学、斯坦福大学及密歇根大学）的上百万图书进行扫描(含 45 亿页文字材料)，实现图书内容的网上搜索及浏览。法国人对此作出了强烈反应。他们担心：一旦这个网上图书馆建成，就可能意味着美国的声音将对人们今后的世界观施加压倒性的影响,而法国曾经创造的优美语言和光辉思想将越来越少为人所知，最终成为被世界遗忘的角落。法总统府在一份声明中写道：“总统先生将与其他欧洲领导人采取行动，以谋求加强在这一领域的协作。当今的世界正在掀起一场知识数字化的革命，拥有独一无二文化遗产的法国和欧洲理应占有一席之地，以便让人们了解欧洲的智慧、历史以及文化遗产”。这也理应成为我们对待中文信息处理的一个新的视角。

（本文发表于《语言文字应用》2005 年第 3 期, 第 38-40 页）