大规模内容计算

白 硕　程学旗　郭 莉　王 斌　余智华　刘 群

中国科学院计算技术研究所 软件研究室 北京 100080
Email: {bai,cxq,guoli,wangbin,yzh,liuqun}@ict.ac.cn

摘　要：随着互联网朝着宽带和廉价方面不断发展，处理大规模信息内容的需求与日俱增。这些需求，来自电信、金融证券、网络安全、重要信息化行业等许多关系国计民生的要害部门和领域。从学术上看，无论从算法上、系统上、还是从深度上，都提出了一系列的研究课题。其中一些课题作为对全人类智慧的挑战，已经纳入年度性的国际测评活动。一些课题由于涉及到国家改革、发展、稳定的大局和学科建设的根本，已经列入一些国家重大科研计划。中科院计算所长期从事大规模内容计算方面的研究开发工作，在这个方向上已经取得了系统而丰富的研究成果，形成了完整的梯队布局。借此机会，我们系统地汇报一下我们在这方面的工作。

关键词：内容计算，信息安全，TREC 会议，数据流，信息过滤，信息检索，自然语言处理，中文分词

Large-scale Content Computing

Bai Shuo　Cheng Xueqi　Guo Li　Wang Bin　Yu Zhihua Liu Qun

Software Division, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080
Email: {bai,cxq,guoli,wangbin,yzh,liuqun}@ict.ac.cn

ABSTRACT: With the development of Internet, the needs for processing large-scale content are coming up. Most of them come from the important national organizations, such as telecommunication and financial departments. From the researcher's point of view, many algorithmic, systemic or deep problems are to be solved. Some of them are challenge to human's intelligence and have been put into international evaluation, and some are very important to the whole country and have been listed in national projects. We have spent long time in large-scale computing and have made some achievements. Some of our works are introduced in this paper.

Keywords: Content Computing, Content Security, TREC, Data Flow, Information Filtering, Information Retrieval, Natural Language Processing, Chinese Segmentation

1 大规模内容计算的内涵和外延

大规模内容计算，是我们近 10 年来的研究工作所围绕的一个中心概念。我们在这里对这个概念略加阐发。

所谓“大规模”，我们是从如下几个方面去理解的：第一是它的处理对象数量规模巨大，基本在 TB 的数量级。第二是它的处理对象往往在时间上具有很强的实时性、在地域上具有很强的分布性。第三是它的需求数量也相对庞大，同时起作用的规则数目可能达到几千到几万。这些特点，是对传统的信息内容处理技术的重大挑战，也决定了相应的处理手段决不是传统方法的简单应用，实验室原型的简单搬家，或者是单兵、散兵作战的组织模式的简单延续所能胜任。它要求一种能够与电信级的运行系统相匹配的、成熟的、高性能的内容处理核心技术和相应的大兵团的研究开发队伍。

所谓“内容”，对我们来说，主要是非结构化的或者半结构化的文本数据。网上的网页、服务器日志、黑客攻击特征数据、网络公告板等等，都属于这个范畴。同关系数据库这类结构化的数据相比，非结构化/半结构化的数据带来了一系列的新问题：第一是不规则的长度使信息的存储和提取缺乏规律可循，不能使用更加有效的提取和存储手段，不便于对信息的“理解”。第二是协议和编码格式的多样化，包括文档本身的文件格式、传输编码格式和网络底层的协议要求的数据报格式等。这些多样化的协议和编码，导致处理的效率和识别的准确性等问题。第三是需求本身涉及到一定的处理深度，比如对话题的跟踪、对倾向性的分析、问答式的服务等等。

所谓“计算”，当然是一种广义的“处理”。这一方面是因为，单纯以获取、检索、过滤、分类、聚类、跟踪、理解、问答等范畴来概括我们的工作，都具有很大的局限性。随着需求的不断扩大，我们的工作也越来越多样化。只有用“计算”的概念才能从更高的高度覆盖我们的研究工作所涉及的处理机制。另一方面，我们不仅研究个别的技术，也开发可供实用的系统。这些系统往往不局限于一种内容处理机制，而是在需求的牵引下，多种内容处理机制的综合集成。再有，我们也需要从计算的角度来审视一些机制之间的内在联系，比如数据库与数据流之间、检索与过滤之间的对偶性等等。

因此，“大规模内容计算”是一个经过深思熟虑后逐渐界定出来的研究方向。它面向规模巨大的、实时性的、结构化程度较弱的数据，采取综合性的计算手段，构造满足具有相当深度和广度的实用化的综合集成系统。

中科院计算所在“大规模内容计算”这一方向上，主要部署了三个层次的研究。系统层，以企事业和国家部门的实际需求为直接目标，构造综合性的内容处理系统。算法层，以系统的需求为背景，结合国际性的测评活动，提出高性能算法。理解层，以追求内容处理的深度为目标，兼顾自然语言处理的传统研究领域，取得了若干创新性的研究成果。

中科院计算所承担的涉及大规模内容计算技术的国家级项目，除企事业支持的应用系统外，还有国家 863、973、自然科学基金和中科院知识创新工程的有关课题。通过这些课题，

我们不仅为应用直接做出了贡献，也为大规模内容计算这个研究方向的形成和相应科学目标的凝练奠定了基础。同时，也为科研团队的成长创造了条件。目前，中科院计算所涉及大规模内容计算的科研团队主要集中在软件研究室，人数已经超过 70 人。通过课题，一批年轻的学术骨干脱颖而出。

2 系统层案例

中科院计算所根据用户的业务需求，开发了多个大规模内容计算实用系统。本文介绍其中的两个系统。限于文章篇幅和有关规定，部分技术细节无法展开论述，敬请谅解。

2.1 针对大规模网络攻击保护的信息内容筛选平台

信息内容筛选平台系统主要为了保障网络安全，可以对大规模广域网网络黑客攻击和病毒入侵等事件及时发现特征并提供响应建议。信息内容筛选的定义是：当被关注的网络流出现符合筛选条件的特征时，及时予以发现并以一定方式告知相关用户。

信息内容筛选平台是一个分布式的大型系统。它可分三个主要的子系统：获取子系统、分析子系统、筛选子系统。

2.1.1 获取子系统 [1, 2]

顾名思义，获取子系统的主要任务是获取网络信息。网络信息的获取，一般采用如下三种方式：

A 主动方式

主动方式即所谓“抓取”方式。其方法为使用一定的“网络机器人”对特定的信息发布源进行轮询，遇到新内容则马上采集过来。这种网络机器人不仅作用于网站/网页，也可以作用于 BBS 和网络聊天室，可以看成是一种支持多协议的“上站机”。

中科院计算所在网络信息的主动获取方面，采取了一些技术措施，使之在安全和效能上达到了新的境界。

首先是获取的安全性。众所周知，获取子系统是按照公开协议、以一个合法客户端的身份登录一个公开的信息系统的。因此，获取子系统不可避免地要暴露在公网上。但是，后台的分析子系统和筛选子系统又需要独立于公网之外以保护自己。我们采用了基于自动隔离开关的“网闸”技术，实现了对后台分析子系统和筛选子系统的安全屏蔽。

其次是获取的效能。信息内容筛选不同于通常的搜索引擎，它具有很强的针对性。我们使用了采样与剪枝相结合的自适应获取技术，对规模较大的网络入口，通过对部分样本的分析来确定是否进行链接的扩展，而不是简单地执行宽度优先的机械搜索。这样就避免了对大

量无关信息的搜索和遍历，提高了信息获取的性能。

B 被动方式

被动方式功能类似大部分 IDS 系统的信息截获。信息获取设备接入网络的特定部位，取得流经该部位的网络流量，随即通过一个高性能协议栈处理机将网络数据包边还原为文本格式边执行扫描和过滤。在本文中，我们不去展开谈论高性能协议栈中与内容处理较少关联的网络技术，而是专注于直接与内容处理相关联的技术。主要包括：多关键词高速字符串匹配算法、扁平协议栈、多表达式匹配算法等等。这些将在第三节中有进一步的介绍。

C 中间人方式

中间人方式是介于主动方式和被动方式之间的一种信息获取方式。它的特点是，通过插入在客户端和服务器端之间的一个透明或半透明的系统，实时对数据流处理，以达到信息特征获取的目的。

2.1.2 分析子系统

分析子系统的任务是对所获取的网络信息按照所设置的需求进行分析，从中筛选出符合筛选条件的部分，交由筛选子系统处理。根据不同任务的需要，内容分析有不同的侧重。

一种是对特定来源的网络信息进行全程记录。这里的主要任务是分析并锁定来源，对后续的内容照单全收，没有分析的必要。此处主要是基于网络数据流的元数据进行半格式化处理。由于网络数据流中的协议类别和特征信息非常多，要想能够精确处理还需要做很多的工作。

一种是根据特定的特征或特征逻辑组合对获取到的网络信息流进行筛选。在面向广域网入侵检测、大规模病毒疫情检测等应用背景的筛选工作中，由于信息量巨大并且实时性强，系统面临很高的性能压力。我们自主设计并实现了由高效的多关键词字符串匹配算法、扁平协议栈技术、模糊匹配技术以及多表达式匹配算法等等组成的内容分析引擎，适应了大信息量、高实时性的要求，较好地完成了任务。

还有一种是根据被分析的对象（比如一个网页或网站）的整体特征的状况是否出现异常（比如，网页/网站是否被黑客所篡改）来进行筛选。这种情况下，性能并不是主要问题，关键是找到正确的判别特征。我们自主设计并实现了一个有效的判别算法，已经用于网页的自动监测。

2.1.3 筛选子系统

筛选子系统根据分析结果生成筛选信息，包括终端上的音频视频手段以及与移动通信设备的互动。这方面的技术与内容计算的核心技术距离较远，不再详述。

2.2 信息内容采集编报平台

信息内容采集编报与信息发布系统，是一个有跨领域共性的应用。它包括收集来自不同来源的信息，对信息进行分类、摘编、概括总结、审核、编纂发布等等一系列过程。网站的

信息发布是最简单的信息编报的例子。新闻机构内部也需要这样的系统，由于新闻机构可能遍布全国甚至全世界，其流程和体系结构非常复杂。

信息采集编报平台涉及到信息获取、内容管理、数据库、工作流、分布式计算等技术的综合运用。大规模内容计算只是其中的部分技术。

中科院计算所设计并实现了一个完整的分布式信息编报系统，该系统目前已经应用到全国很多省市的相关应用单位。在信息获取和内容管理方面，采用了属于大规模内容计算范畴的若干新技术，比如自适应剪枝和个性化推送技术。

3 算法层案例

算法是大规模内容计算领域核心竞争力的最主要的体现。中科院计算所在大规模内容计算的算法方面取得了许多具体的成果。其中一些算法用在了应用系统的核心模块当中，一些算法在国际性的测评活动中取得了好成绩。本文中，我们将介绍其中比较有代表性的几项工作。

3.1 多关键词高速字符串匹配算法 [3]

入侵检测、病毒检测、个性化主动服务等许多应用都离不开对特定关键词是否出现在特定文档中的判断。多项需求中的多个用户都需要从同一个网络数据流中找到自己所需要的东西。我们不能设想为这么多项需求中的每一个需求增加一遍对整个数据流的扫描，因为那意味着巨大的浪费，也使实时性成为不可能。大数据量、实时性的要求，迫使我们寻求更加高效的算法。一个核心的制约就是：一次扫描解决问题。

如果暂时不考虑不同网络层次的协议栈实现问题，把问题暂时抽象到一个层次，这就是一个多关键词字符串匹配的问题。

中科院计算所提出了多个算法来解决这个问题。其中，一个最好的算法 IntMatch 是这样一个思路：把数据流和关键词都编码为长整数序列，根据以模除法为基础的散列机制来排除不匹配的部分，最大可能地借助 CPU 的硬件流水功能实现并行处理，同时最大可能地根据关键词的特点实现跳跃式扫描。实验表明，这个算法是在目前世界上已经发表的多关键词匹配算法中，匹配效率是最高的。

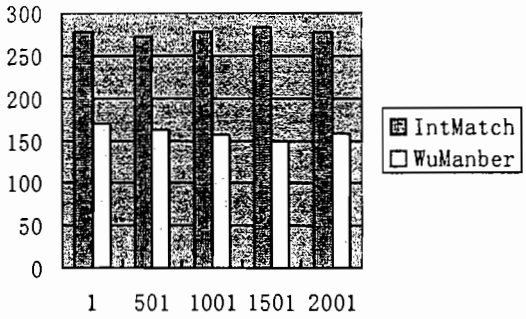

图 1：中科院计算所提出的 IntMatch 与此前世界上最好的多关键词匹配算法 WuManber 的性能比较。横轴：关键词个数，纵轴：速度（MB/s）

3.2 扁平协议栈

实际上，任何文本在网络上最终被获取到的形态，都是经过多个层次的编码的。如果按步就班地一个层次一个层次地解码、扫描、匹配，必将导致多次扫描，使得实时性的需求无法得到满足。为了解决这个问题，中科院计算所自主涉及并实现了基于扁平协议栈的字符串匹配算法。该算法的实质是在考虑到逐层编码的现实条件下，打破层次界限，一步到位地实现必要的解码和匹配。

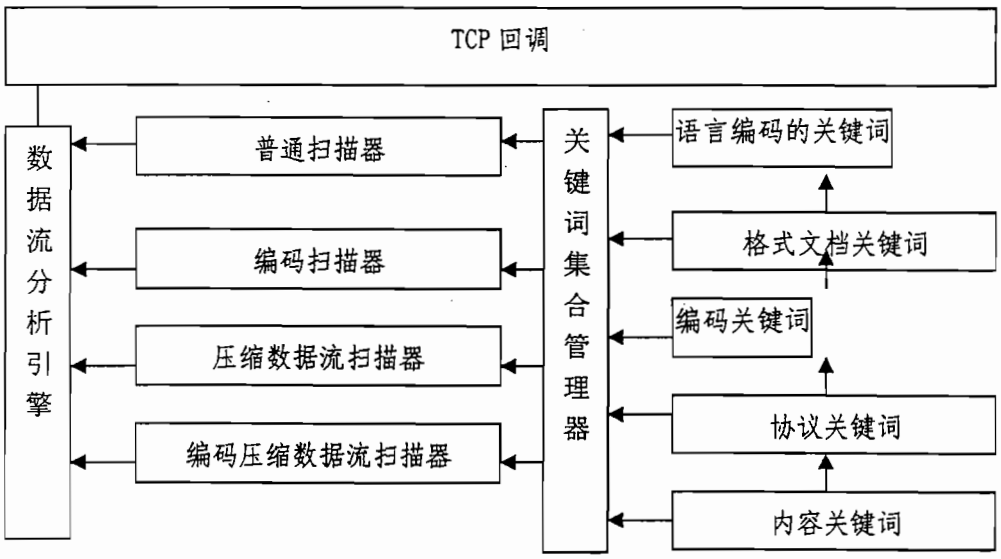

图 2 扁平协议分析的数据流执行自动机结构图

3.3 多表达式匹配算法

对大规模内容计算的实际的应用需求往往是有内在逻辑结构的。就是说，有多个用户，每个用户可能有多项需求，每项需求涉及到若干个关键词的特定逻辑组合。这样，一个支持多用户、多需求的匹配引擎，最核心的部分可能就是一个多表达式匹配算法。

中科院计算所提出的算法，以自动机为基础，改进了经典计数算法（predicate counting algorithm）中不适合数据流环境的数据结构，可以在线性时间内完成组合关键词匹配。组合关键词匹配算法还可以广泛地应用于基于调用序列的入侵检测系统中。

3.4 参加 TREC 之自适应过滤测评任务的算法 [4, 5]

2001-2002 年，中科院计算所相继参加了 TREC-10 和 TREC-11 的自适应过滤任务评测。

所谓自适应过滤，就是要在一个提供一定相关反馈信息的环境下，尽快捕捉用户的兴趣，以求在后续的筛选过程中取得逐步贴近用户需求的过滤结果。跟 TREC-10 过滤任务相比，TREC-11 过滤任务的难度加大了。给出的相关反馈信息更加不充分。因此，从这种小样本的反馈中学习到准确的、稳定的用户兴趣就更加不容易。

面对难度增大的任务，我们重点从需求扩展、小样本自我复制和自我增强等方面着手，改进用户个性化需求的获取效果，并改进了特征选择算法。在参数的调整方面，我们采用了新的优化算法。

在 TREC-11 自适应过滤任务的最终评测中，我们取得了两项指标（T11SU 和 T11F）均排名第一的优异成绩，并且遥遥领先其他参加者。表 1 是我们提交的过滤结果。表 2 是我们的结果跟第二名（KerMIT）、第三名（CMU）结果的比较。容易看出，我们的过滤系统在性能方面的优势是很明显的。这说明我们的过滤系统对于不同的过滤环境具有良好的适应能力。

Run ID	MeanT11SU	MeanT11F	Precision	Recall	Retrieved	Relevant
ICTAdaFT11Ua	0.405	0.244	0.310	0.197	4166	2532
ICTAdaFT11Ub	0.403	0.245	0.311	0.199	4261	2543
ICTAdaFT11Uc	0.403	0.242	0.307	0.200	4214	2546

表 1 中科院计算所提交的结果在 TREC-11 自适应过滤任务中的评测结果

Run ID	MeanT11SU	MeanT11F	Precision	Recall	Retrieved	Relevant
ICTAdaFT11Ua	**0.405**	**0.244**	**0.310**	**0.197**	**4166**	**2532**
KerMITT11af1	0.390	0.214	0.334	0.140	2876	2010
CMUDIRUml	0.369	0.222	0.279	0.157	3077	1817

表 2 中科院计算所提交的结果在 TREC-11 自适应过滤任务中与第二、三名的结果比较

下面的图 3 和 4 分别是 TREC-11 自适应过滤任务的所有参加单位在 T11SU 和 T11F 两个主要评测指标上的完全评测结果。其中，前三个是中科院计算所提交的结果。

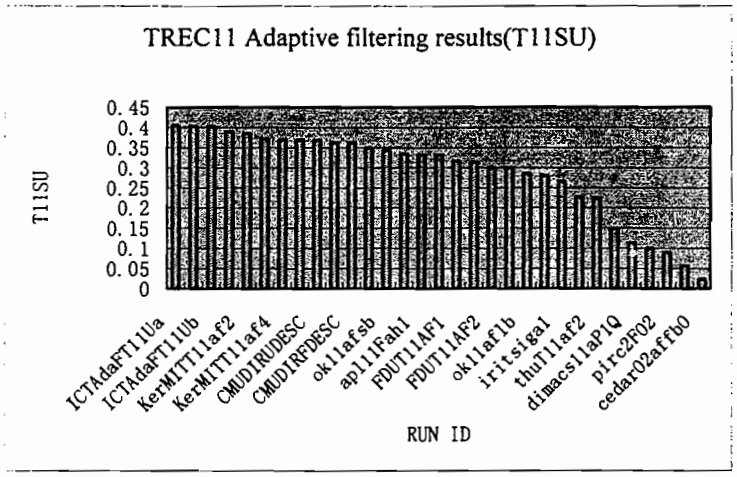

图 3 TREC-11 自适应过滤任务的 T11SU 完全评测结果

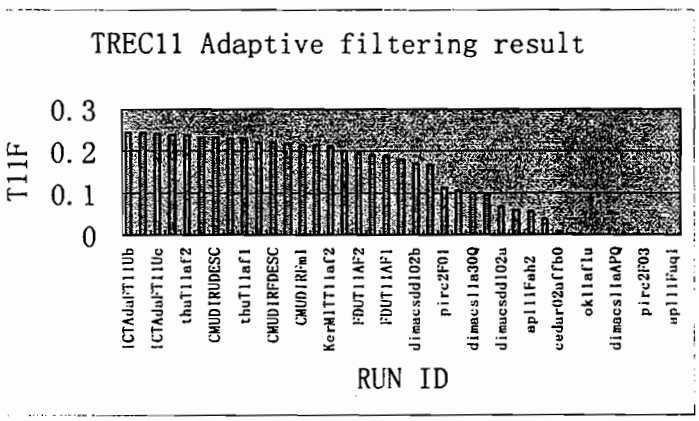

图 4 TREC-11 自适应过滤任务的 T11F 完全评测结果

3.5 参加 TREC 之 Web 检索测评任务的算法[5]

中科院计算所参加的另一项 TREC-11 测评任务是 Web 检索。我们参加了其中的两个子任务：一个是命名页面发现（Named Page Finding），另一个是主题浓缩（Topic Distillation）。在命名页面发现子任务中，我们采用多元特征综合分析的方法，同时考虑了内容相关性得分、链接文本和文档结构等因素；在主题浓缩子任务中，我们考虑了内容相关性得分、链接分析、

基于 URL 长度的重新计算权重等策略。在参数调整过程中，我们使用了基于统计分析的方法。我们提交的运行结果给出了较为令人满意的性能表现。

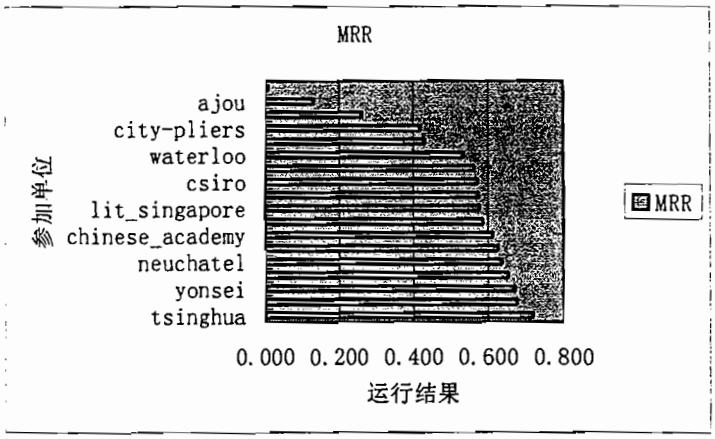

图 5 TREC 11 Web Track – Named Page Finding – MRR

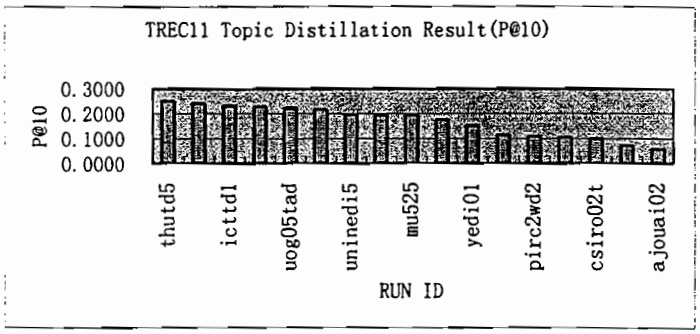

图 6 TREC-11 Web Track – Topic Distillation – P@10

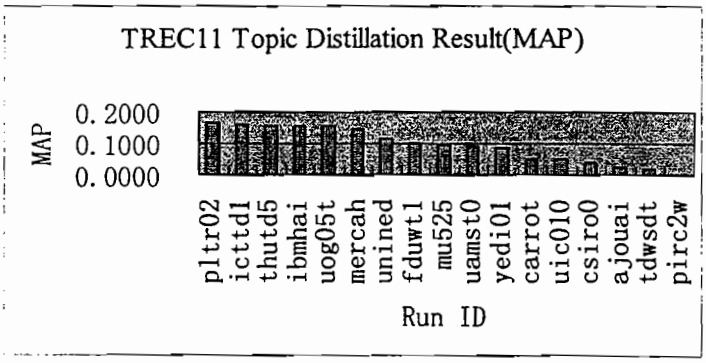

图 7 TREC-11 Web Track – Topic Distillation – Mean Average Precision

4 理解层案例

中科院计算所对于大规模内容计算的新生长点和向一定的理解深度发展历来非常重视。我们不仅根据实际需要大胆尝试新的思路，而且部署了面向自然语言处理的基础研究。我们认为，随着可以使用的计算资源的规模越来越大，内容计算既满足实用上的性能需要、又达到一定的理解深度是完全可能的。因此，我们一方面鼓励在这个方向上的创新思维，另一方面更鼓励新的想法能够在大规模内容计算中得到检验和验证。话题发现与跟踪和倾向性分析就是这方面的成功案例。在面向自然语言处理的基础研究当中，我们在分词、句法分析、人机对话和资源建设等方面都取得了很有特色的成绩。

4.1 话题发现与跟踪

网络上的信息，尤其是新闻信息，呈现出一种结构，这就是话题-线索结构。话题体现具有共同中心特征的文档的自然聚类（某一群文档的“向心性”），线索体现话题随时间的展开（时序性）。人们对信息的筛选，也在很大程度上循着话题-线索结构来进行。因此，大规模内容计算除了针对传统的向量空间结构和链接结构之外，也可以选择话题-线索结构作为一种计算对象，这就是国际上所说的 TDT（Topic Discovery and Tracking）[6]，即话题发现与跟踪。话题发现与跟踪在商业上可以发展成一种嵌在搜索引擎中的崭新的信息服务模式，就是说以话题-线索结构为依托向用户进行有针对性的推送。要做到精确的推送，必须有精确的话题发现与跟踪能力。目前的技术做到全自动的发现跟踪是有相当难度的。我们采用一种人机交互的方式来从用户的行为中产生相关的话题-线索结构，也就是说，通过自动聚类产生一些初始的“团”，然后再通过用户行为的反馈对初始的“团”进行修正，达到比较满意的中心特征提取和线索预报。我们承担了一个国家专项研究任务，按照这种技术路线来帮助有关部门跟踪特定的信息。

4.2 倾向性分析

对于文本的检索、过滤和分类来说，最熟悉不过的就是向量空间模型了。但是，向量空间模型有一个难以克服的问题，就是各个特征分量之间是一种“扁平”的关系，体现不出“层次感”。一个后果就是，用这种模型去处理针对某些实体的不同态度所形成的类别，往往效果不佳。比如对某支股票的涨与跌的预测、对某一个国际性事件（如美伊战争）的支持与反对等等。我们把这类问题称为“倾向性分析”。

对于“倾向性分析”来说，我们关心的不是词在文本当中的出现状况，而是文本中围绕着一些词的另一些词的出现状况。这样一来，特征的表达就要打破传统的向量空间模型那种“扁平”的结构，而出现一种类似“张量”的结构：我们把词项分成“实体词项”和“倾向词项”。对于一个特定的倾向性分析任务，所涉及的实体词项是一个不大的封闭集合。一个文本，表示为所有相关的实体词项周围一个固定邻域范围内的倾向词项的出现状况的统计和（中科院计算所鲁松博士的研究成果表明，围绕一个词的前 8 个词、后 9 个词可以提供 85%以上的信息量）。利用这种“张量空间模型”，我们在倾向性分析中取得了意想不到的好结果——二值分类开放测试的结果高达 95%以上。

4.3 中文分词 [7]

分词是面向汉语的自然语言处理领域的传统问题，也是许多大规模内容计算任务中不可缺少的共性模块。中科院计算所在分词算法方面投入了很大的精力，使用多层隐马尔可夫模型和 N-best 搜索算法，得到了比较理想的效果，在人名、地名、机构名等未登录词的识别上效果尤其显著。这方面的工作发表在一系列的国际国内杂志和会议上，受到普遍的肯定和重视。我们根据这些技术构造的分词系统 ICTCLAS，在共有 13 支队伍参加的第一届国际分词测评比赛中，参加了 6 项测试，在其中的 4 项基于简体语料库的测试中，CTB closed 和 PK closed 两项测试中名列第一，PK open 测试中名列第二。该分词系统也成为中科院计算所首批向社会公开的研究成果之一，已经成为中科院计算所自然语言处理开放平台(www.nlp.org.cn)上最热门下载的软件。

4.4 概率型句法分析算法 [8]

自然语言的情况是极其复杂的。合语法的，不一定满足语义搭配关系。满足语义搭配关系的，又不见得是一种“常态”的说法。因此，针对一种复杂的自然语言（比如汉语），按照上下文无关语法写出来的规则集合，在一个确定性的句法分析器上，往往会得到非常多的分析结果，而其中我们真正想要的结果只是极少数，大多数属于“伪歧义”，而且这种伪歧义并不会随着句法规则的逐步求精而趋于消失。剩下的问题，只能用概率的方法解决，依靠大量成功的分析结果即所谓“树库”来排除非“常态”的分析结果。我们在自创的“角色反演算法”中引入概率因素，大大提高了分析的正确率。这个系统在我们的自然语言处理开放平台上有演示版本，欢迎大家测试。

但是，我们所依托的树库是美国的宾州树库，它的某些先天性缺憾也影响了系统无法得到更好的发挥。我们借此机会，也呼吁国内同仁，重视树库的建设，吸收中文句法体系研究的最新研究成果来完善树库的顶层设计，树库建设与中文句法分析器调试环境的建设同步进行。我们无偿地牵头建设自然语言处理开放平台，其中一个目的也正是想要集中所有学者的智慧，积累所有同仁的劳动成果，建立这样一个开放的、合作的测试环境。

4.5 模式推理 [9]

问答系统缺乏推理能力，推理系统缺乏自然语言理解能力，这是一个老问题了。正是这个问题困扰着大型知识库系统的建设，也使花费巨大的人力物力建立起来的知识库系统难以面向公众开展达到一定质量的知识服务。中科院计算所提出了“模式推理”的思想，以自然文本形态作为知识的存储形态，免去了知识形式化的环节；同时，增加了语句模板之间的推理机制，拓宽了浅层知识表示的推理深度和排除歧义的能力。我们以“红楼梦”人物关系为背景，做成了一个试验问答系统。系统同时表现出一定程度的问题理解能力和外在推理能力，但却没有一个人为形式化的知识库。一切输入并存储的知识都是自然语言语句。目前，我们正在对模式的合一和联立合一的算法以及推理控制的算法进行进一步的研究，争取在更大的规模上实现更加复杂的推理机制。

5 结语

中科院计算所在大规模内容计算方向上的研究工作，是在统一的战略布局下，面向国家的重大关键需求，基于多年研究工作的深厚积累，有层次地来开展的。我们感谢用户单位的的支持和信任，也感谢广大同行们的友好合作。我们认为，大规模内容计算将随着信息网络技术的进一步发展，从实际应用中汲取更多的营养，逐渐形成自己的相对独立的领域、目标、方法和理论体系，更好地解决实际应用中遇到的问题。我们所接触到的实际应用领域也正在为信息检索、自然语言处理等传统研究领域带来新的活力。这种活力值得我们高度重视。我们希望有更多的同行同我们一道开拓大规模内容计算这个新兴的方向，努力取得更多更好的成果来，为信息安全、国民经济发展和学科建设做出贡献。

参 考 文 献

[1] 余智华. WWW站点的分析与分类. 硕士论文. 1999.

[2] 李盛韬. 基于主题的Web信息采集技术研究. 硕士论文. 2002.

[3] 谭建龙. 串匹配算法及其在网络内容分析中的应用. 博士论文. 2003.

[4] Bin Wang, Hongbo Xu, Zhifeng Yang, Yue Liu, Xueqi Cheng, Dongbo Bu, Shuo Bai. TREC-10 Experiments at CAS-ICT: Filtering, Web and QA. In The Tenth Text REtrieval Conference (TREC 10). page 109, 2001.11, Gaithersburg, MD, USA.

[5] Hongbo Xu, Zhifeng Yang, Bin Wang, Bin Liu, Jun Cheng, Yue Liu, Zhe Yang, Xueqi Cheng, Shuo Bai. TREC-11 Experiments at CAS-ICT: Filtering and Web. In The Eleventh Text REtrieval Conference (TREC 11).

2002.11, Gaithersburg, MD, USA.

[6] Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation. Wayne. C., Language Resources and Evaluation Conference (LREC) 2000, pages 1487-1494.

[7] Kevin Zhang, Qun Liu, Hao Zhang, Xueqi Cheng. Automatic Recognition of Chinese Unknown Words Based on Role Tagging. In the proceedings of COLING 2002 Post-Conference Workshop, First SigHan Workshop, Taipei, September 2002

[8] 白硕，张浩. 角色反演算法.软件学报.2003(3):.328-333.

[9] 王树西.一个人物关系问答的专家系统.第七届中国人工智能联合学术会议论文集. 2003:31~36.